DOCUMENT RESUME

ED 415 751 HE 030 895

AUTHOR Kitchener, Karen S.
TITLE Assessing Reflective Thinking within Curricular Contexts.
INSTITUTION Denver Univ., CO.
SPONS AGENCY Fund for the Improvement of Postsecondary Education (ED), Washington, DC.
PUB DATE 1994-02-28
NOTE 82p.
CONTRACT P116B00926
PUB TYPE Guides - Classroom - Teacher (052) -- Reports - Descriptive (141) -- Tests/Questionnaires (160)
EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Adult Development; Cognitive Development; Cognitive Tests; *College Curriculum; Conceptual Tempo; Consultation Programs; *Critical Thinking; Educational Psychology; *Faculty Development; Higher Education; *Instructional Improvement; Models; Theories; Thinking Skills; Workshops
IDENTIFIERS Bowling Green State University OH; *Reflective Thinking; *University of Denver CO; University of Missouri Columbia

ABSTRACT

This 42-month project at the University of Denver (Colorado) developed materials and faculty development activities concerning application of Kitchener and King's reflective judgment theory. Testing and implementation were at the University of Denver and Bowling Green State University (Ohio) and the University of Missouri (Columbia). The model describes the development of one aspect of critical thinking--the process by which adults become better able to make decisions about complex ill-structured problems. A paper and pencil measure, the reflective thinking appraisal, was devised and tested, and an accompanying technical manual was written. Three-day workshops and consultation with 18 faculty in various disciplines led to revision of courses based on the reflective judgment model and principles of educational psychology. A manual was developed to help faculty apply the model. Evaluation of reflective judgment in pilot studies indicated significant differences in scores between younger, less educated students and older, more educated students, with reliability over the entire age range estimated at .79. Questionnaires and interviews with participating faculty revealed a greater understanding of students' need for support in developing critical thinking skills and increased awareness of ill-structured problems in their disciplines. Appended are the reflective thinking measure, the accompanying technical manual, and the workshop evaluation form. (DB)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 415 751

# Assessing Reflective Thinking Within Curricular Contexts

Grantee Organization:

University of Denver
College of Education
Denver, CO 80208

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
□ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Grant Number:

P116B00926

Project Dates:

Starting Date: September 1, 1990
Ending Date: February 28, 1994
Number of Months: 42

Project Director:

Karen S. Kitchener
College of Education
University of Denver
Denver, CO 80208
(303) 871-2480

FISPE Program Officer: Preston Forbes

Grant Award:

| | |
|---|---|
| Year 1 | $ 96,414 |
| Year 2 | 114,174 |
| Year 3 | 83,664 |
| Total | $ 294,252 |

BEST COPY AVAILABLE

FIPSE Project P116B00926 Paragraph Summary

HE 030 895

FIPSE FINAL REPORT--Project #P116B00926

Assessing Reflective Thinking Within Curricular Contexts


Page

Cover Sheet . . . . . 1

Paragraph Summary . . . . . 2

Executive Summary . . . . . 3

Project Overview . . . . . 6

Purpose . . . . . 6

Background and Origins . . . . . 9

Project Description

Goals 1 and 2: Assessment Instrument . . . . . 11

Goals 3 and 4: Faculty Consultation . . . . . 14

Results

Goals 1 and 2: Assessment Instrument . . . . . 15

Goals 3 and 4: Faculty Consultation

University of Denver . . . . . 16

Bowling Green State University . . . . . 18

Summary . . . . . 19

Next Steps . . . . . 20

Summary and Conclusions . . . . . 22

Tables and Figure

Appendices

A--*Reflective Thinking Appraisal*

B--*Technical Manual to Accompany the Reflective Thinking Appraisal*

C--Sample Evaluation Form From Faculty Workshop

D--*Developing Reflective Judgment in the Classroom: A Manual for Faculty*

E--Information for FIPSE

"Assessing Reflective Thinking Within Curricular Contexts"

This project was based on Kitchener and King's Reflective Judgment theory. The model describes the development of one aspect of critical thinking, the process by which adults become better able to make decisions about complex problems that do not have clear-cut right and wrong answers. A paper and pencil measure of Reflective Judgment was devised and tested, and an accompanying technical manual was written. In addition, through consultation with faculty in a variety of disciplines, courses were revised based on the Reflective Judgment model and principles of educational psychology. Finally, a manual was written to help other faculty apply the Reflective Judgment model.

Karen S. Kitchener
College of Education
University of Denver
Denver, CO 80208
(303) 871-2480

Project Products

*Developing Reflective Judgment in the Classroom: A Manual for Faculty*

*Reflective Thinking Appraisal*

*Technical Manual to Accompany the Reflective Thinking Appraisal*

Executive Summary--FIPSE Project P116B00926

"Assessing Reflective Thinking Within Curricular Contexts"

University of Denver
College of Education
Denver, CO 80208

Karen S. Kitchener
(303) 871-2480

**Project Overview**

This project focused on devising a paper and pencil measure of one aspect of critical thinking and on helping faculty improve the critical thinking of their students. The theoretical basis for the project was the Reflective Judgment model, developed by Karen Kitchener, University of Denver (DU), and Patricia King, Bowling Green State University (BGSU). The model is an empirically validated tool for understanding how students develop the ability to address complex problems for which no single, absolutely correct solution can be determined (called ill-structured problems). A two-problem paper and pencil measure of Reflective Judgment was developed and tested with the assistance of Phillip Wood, University of Missouri--Columbia (UM-C). Its development is continuing in collaboration with educators and researchers at other institutions. Consultation with small groups of faculty at DU and BGSU and presentations to a group of faculty from UM-C were aimed at helping faculty better understand the developmental needs of their students and be more deliberate in their efforts to encourage the development of reflective thinking. As a result of this aspect of the project, a manual was written for use by other interested faculty. Both the instrument development and the faculty consultation will continue after the end of the FIPSE funding period.

**Purpose**

Four purposes were addressed in this three year project:

1. To develop an objective measure of Reflective Judgment that could be used by faculty and administrators to evaluate the nature of students' reasoning about ill-structured problems.
2. To adapt the Reflective Judgment Interview to the specific content of three disciplines: business, chemistry, and psychology.
3. To use data on Reflective Judgment already collected to consult with faculty about how to adapt their instruction to the developmental characteristics of students.
4. To develop materials that would allow faculty at other sites to use the objective measure and to apply the Reflective Judgment model to their teaching.

**Background and Origins**

The project grew directly out of the theoretical and research work of Drs. Patricia King and Karen Kitchener. The project team consisted of Drs. King and Kitchener, Dr. Cindy Lynch, and Dr. Phillip Wood. Three institutions were represented by these educators; the institutions

varied in size, geographic setting, student characteristics, and history/mission. These differences contributed to the richness of the project, but also created some unanticipated administrative difficulties.

The Reflective Judgment model had come to the attention of educators at a variety of institutions through professional publications as well as more popular ones such as *Time* and *Omni*. These educators were asking for a way to measure reflective thinking that was more time and cost efficient than the Reflective Judgment Interview, which has been used to empirically validate the model. Similarly, at the three primary institutions served by the project, concern was increasing about assessing outcomes of higher education. The project team envisioned the Reflective Judgment model as a tool for improving teaching to encourage one aspect of critical thinking and for devising effective assessment strategies.

**Project Description**

Although the first year of the grant period was spent attempting to develop a computerized measure of Reflective Judgment, data indicated that these attempts were not very successful, with reliability estimates falling well below acceptable standards. After repeated revisions and piloting of the computerized measure, the project team decided to move away from that format to a paper and pencil format. The resulting assessment instrument is called the Reflective Thinking Appraisal (RTA), and it currently includes two problems: one with historical content about how the pyramids were built, and one with scientific content about the safety of artificial sweeteners in foods. Ill-structured problems in chemistry, business, and psychology also were written and successfully piloted in the interview format. These stand ready for use in later development of the RTA.

Prior to commencing the project, three faculty who had experience using the Reflective Judgment model in their teaching of chemistry, psychology, and English had been identified. These consultants met with the project team to share their experiences, concerns, and advice. The information gleaned from this meeting was used to design 3-day workshops for faculty interested in improving the reflective thinking of their students. Small groups of faculty from the College of Business Administration at DU and from a variety of disciplines at BGSU participated in three-day workshops on their respective campuses and revised at least one course they taught in the subsequent year based on the information they received in the workshop. Consultation with these participating faculty continued in the small groups and individually during subsequent academic terms.

**Results**

The current version of the RTA is a viable measure for investigating one aspect of critical thinking, how students understand and solve ill-structured problems. It takes between 40 minutes and an hour for students to complete the measure. Data from pilot studies of the instrument indicate significant differences in scores between younger, less educated students and older, more educated students, with reliability over the entire age range estimated to be .79. See the

*Technical Manual to Accompany the Reflective Thinking Appraisal* for a complete description of the measure and related data.

Most of the faculty participants in the project indicated that the notion of providing developmentally appropriate support as they teach students to become better at making reflective judgments was a key positive outcome for them. For example, attending to sequence and structure of course content provides both intellectual and emotional support for students, and attending to the emotional needs of students helps them be more receptive to the challenges they face in classes designed to encourage reflective thinking. *Developing Reflective Judgment in the Classroom: A Manual for Faculty* was written as a result of the team's consultation efforts. It can be used by faculty as a tool for modifying their courses to more deliberately encourage reflective thinking.

**Summary and Conclusions**

This FIPSE project resulted in a viable version of the Reflective Thinking Appraisal, a paper and pencil instrument designed to assess how students think about ill-structured problems. A technical manual to accompany this assessment instrument also is available. The project also helped small groups of faculty at the University of Denver, Bowling Green State University, and the University of Missouri--Columbia to redesign courses so that students are deliberately encouraged to improve their reflective thinking skills. A manual for use by faculty was written based on the project team's consultation with participating faculty.

In addition to the assessment instrument, the technical manual, and the faculty manual, the project team gained two generally valuable insights: First, developing sound, well-validated instruments to assess critical thinking is a long-term project. Even though the instrument development efforts in this project were founded on a valid theoretical model and a much used interview assessment strategy, three and a half years was enough time for only the initial stages of the instrument development. Second, efforts at truly successful faculty development require sustained interest and work on the part of both the participating faculty and the consultants. A three-day workshop can only put the process of curricular change in motion. Continued contact among faculty groups and between faculty and consultants is an important part of the long-term work of curricular adaptation to encourage reflective thinking.

The project set the stage for continuing instrument development and faculty consultation. For example, with an National Science Foundation grant at the University of Denver, faculty are applying the reflective judgment model to the teaching of core natural science courses, and the RTA is being used to assess one outcome of that project. A currently funded FIPSE project at Vanderbilt University is piloting the use of the RTA as an outcome measure for students participating in service-learning activities. Efforts are being made to inform other researchers and educators about the results of the project and to include other institutions in this on-going work.

## PROJECT OVERVIEW

This project focused on devising a paper and pencil measure of one aspect of critical thinking and on helping faculty improve the critical thinking of their students. The theoretical basis for the project was the Reflective Judgment model, developed by Karen Kitchener, University of Denver (DU), and Patricia King, Bowling Green State University (BGSU). The model is an empirically validated tool for understanding how students develop the ability to address complex problems for which no single, absolutely correct solution can be determined (called ill-structured problems). A two-problem paper and pencil measure of Reflective Judgment was developed and tested with the assistance of Phillip Wood, University of Missouri--Columbia (UM-C). Its development is continuing in collaboration with educators and researchers at other institutions. Consultation with small groups of faculty at DU and BGSU and presentations to a group of faculty from UM-C were aimed at helping faculty better understand the developmental needs of their students and be more deliberate in their efforts to encourage the development of reflective thinking. As a result of this aspect of the project, a manual was written for use by other interested faculty. Both the instrument development and the faculty consultation will continue after the end of the FIPSE funding period.

## PURPOSE

One consistently identified goal of postsecondary education is the development of students' abilities to reason clearly about and solve complex problems in the face of uncertainty. However, those charged with accomplishing this goal and documenting student outcomes in this arena have been hampered by a lack of understanding about the process through which these abilities develop and by the absence of accessible, conceptually grounded, and well-validated instruments to assess students' progress. The Reflective Judgment model provides an empirically validated tool for understanding how students develop the ability to address complex problems, and the Reflective Judgment Interview provides a starting point for thinking about and designing

tools to assess students' reasoning about problems for which no single, absolutely correct solution can be determined (called ill-structured problems).

The original purposes of this project were:

1. to develop a computerized objective assessment instrument for the Reflective Judgment development model;

2. to adapt that instrument to the specific content of at least two disciplines;

3. to teach faculty to use the assessment information to adapt their instruction to the developmental characteristics of their students; and

4. to prepare documents that allow the assessment procedures and teaching materials to be used by instructors not included in the original project.

The form of these purposes changed somewhat over the course of the project, but the emphasis on developing assessment tools and strategies and on working with faculty remained paramount. The modified purposes are described in the following paragraphs.

1. The first purpose of the project was **to develop an objective measure of Reflective Judgment that could be used by faculty and administrators to evaluate the nature of students' reasoning about ill-structured problems**. The measure needed to be one that also was reasonably reliable and remained meaningfully related to the characteristics of the Reflective Judgment model that it was intended to assess. Despite beginning with a validated model and an interview measure that was psychometrically sound, moving to a computerized assessment was too large of a step to complete in three years. This time consuming and complex task took more rounds of data gathering and more intense problem solving as a team than initially anticipated.

Two administrative difficulties made the completion of this part of the project difficult. Both were tied to the fact that the project was located at three different sites. The first had to do with negotiating subcontracts that were acceptable to all institutions. This was particularly problematic because one institution was concerned about patent rights to any computer software that was developed. It took almost a year to negotiate that subcontract, and this delayed work on the measure at that site. The second problem arose because it took more time to conceptualize

and revise the measure together as a working team than originally anticipated. Consequently, we were under budgeted for team meetings and had to creatively find opportunities to work together (e.g., at conferences) at our own expense or at the expense of our institutions.

2. The second purpose of the grant was **to adapt the Reflective Judgment Interview to the specific content of three disciplines, business, chemistry, and psychology**, in anticipation of the time when the paper and pencil measure could be adapted to the content of different disciplines. This goal was accomplished with few impediments. Faculty and students seemed to welcome discipline-specific questions. The development of the psychology problems provided the basis for a master's thesis at the UM-C.

3. The third purpose of the project was **to use data on Reflective Judgment already collected at each of our institutions to consult with faculty about how to adapt their instruction to the developmental characteristics of students**. Although considerable contact work had been done with faculty and administrators prior to initiating the project, different obstacles arose on each site. At the UM-C, the director and a key faculty member in the campus writing intensive program left the institution. We had planned to use students from the program to pilot the new Reflective Judgment measure and to focus our third year UM-C faculty consultation efforts with the faculty of the program. Because the program was without leadership for an extended period of time, we had to find a different group of students to sample and faculty with whom to consult.

At DU while the dean and individual faculty members in the College of Business Administration were enthusiastic about spending time on improving instruction, department chairpersons sometimes were not as supportive and questioned the time faculty members were devoting to the project. At BGSU the original plan was to work with faculty in the Department of Chemistry. While some faculty were highly enthusiastic, others questioned the time commitment to teaching required by the project as well as the philosophy of science that provided the foundation for it (specifically, whether ill-structured problems exist in chemistry). In the end, this precluded us from relying exclusively on the BGSU chemistry department for participants in

the faculty consultation aspect of this project at BGSU. Solving each of these problems took intensive consultation time from the team leader at each site.

4. The last purpose of the grant remained as initially described: **to develop materials that would allow faculty at other sites to use the objective measure and to apply the Reflective Judgment model to their teaching**.

## BACKGROUND AND ORIGINS

This project grew directly out of the prior theoretical and empirical work of Dr. Karen Kitchener at DU and Dr. Patricia King at BGSU. Drs. Kitchener and King had developed the Reflective Judgment model and Reflective Judgment Interview over the prior 15 years. Dr. Phil Wood at UM-C and Dr. Cindy Lynch at DU had been closely tied to earlier efforts to assess reflective judgment, and Dr. Wood in particular had statistical and computer expertise that were necessary prerequisites to develop a measure that was psychometrically sound. Dr. Lynch had prior experience administering and coordinating grants. The complimentary nature of the expertise of these team members was essential to the progress made on the objective measure, which was called the Reflective Thinking Appraisal (RTA).

In addition, the core team had identified several faculty from other institutions who had expertise in applying the Reflective Judgment model as well as other developmental constructs to the teaching/learning process. These included Dr. David Finster from Wittenberg University in chemistry , Dr. Barry Kroll from Indiana University in English, and Dr. Katherine Nevins from Bethel College (St. Paul, MN) in psychology. These experienced faculty were willing to consult with the project team about applying the Reflective Judgment model in the classroom.

The Reflective Judgment model had come to the attention of educators at a variety of institutions through professional publications as well as more popular ones such as *Time* and *Omni*. These educators were asking for a way to measure Reflective Judgment that was more time and cost efficient than the Reflective Judgment Interview. In fact, in the course of this project we have had over 30 requests for information about using the RTA at different

institutions. Similarly, at our own institutions concern was increasing about assessing outcomes of higher education. In addition, because of the attention given to the Reflective Judgment model in the popular press, faculty were approaching the project directors to work with them individually on improving the reflective thinking of their students. Because of the interest at the UM-C in outcomes assessment Dr. Wood had received university funding to begin work on a computerized version of the Reflective Judgment Interview prior to beginning the grant.

The three institutions varied in size, geographic setting, student characteristics, and history/mission. The differences in these institutions and particularly in the departments/colleges which were initially targeted for our faculty consultation efforts made important differences in the outcome of this portion of our efforts. Most notably the College of Business Administration at the University of Denver had received a $10,000,000 grant to improve instruction in business. Areas of interest included creativity and moral judgment. Because reflective judgment was perceived to relate to both of these goals, business faculty were "primed" to be interested in our project and to continue with it after FIPSE funding ended. Prior to initiating the FIPSE grant, Dr. Kitchener had been invited by the faculty improvement committee to do a half-day workshop for the entire business faculty on using the model in teaching prior to initiating the FIPSE grant. In addition, the business faculty was large consisting of approximately 60 members; thus, there was no problem in recruiting faculty for a 3-day workshop.

By contrast, at BGSU Dr. King had been approached by two faculty members in chemistry about helping them develop reflective thinking in their classes. Because Dr. King had already been contacted by Dr. David Finster from Wittenberg University about his work applying reflective judgment to teaching and because he agreed to help with a workshop at BGSU, the chemistry department appeared to be an appropriate place to use the Reflective Judgment model in the classroom. Two forces mitigated against this occurring. First, the Chemistry Department at BGSU is research grant driven and relatively small (16 members). Consequently, several faculty were unwilling to devote large amounts of time to this project. Second, because of the size of the department there were not enough willing members to devote an entire workshop to

improving teaching in chemistry. At about the same time, Dr. King had been approached by several faculty in other departments who were willing to make such a time commitment, and thus, the workshop at BGSU was composed of faculty from several disciplines.

Last, several forces were at work at UM-C. As already noted, the original plans to work with the Writing Intensive Program were thwarted when a key faculty member and administrative leader left the University. However, UM-C faculty from a variety of disciplines have developed a commitment to teaching that is expressed each spring in a week long retreat which has come to be known as Wakonse. This faculty group has on-going meetings during the academic year and is supported with institutional funding. Dr. Wood had participated in Wakonse in prior years and had been invited to speak on reflective judgment at one of the on-going meetings during the academic year. As a result, the entire team with our three outside consultants were invited to make presentations to faculty in a variety of disciplines from the University of Missouri and other participating institutions at the 1993 Wakonse spring retreat.

## PROJECT DESCRIPTION

**Goals 1 and 2: Assessment Instrument.** As noted earlier, the details of the original goals were modified as we progressed through the project. The most significant changes were made in regard to development of an assessment instrument. Our timeline and goals for instrument development were too ambitious. The development of sound, well-validated assessment instruments generally takes years of work, even when the researchers have a well-validated theory from which to work, as was the case in this project. During the first year, discipline-specific, ill-structured problems were written with the input of faculty who would eventually participate in the faculty workshops sponsored by the project. These problems were pilot tested in the interview format.

During the spring and summer prior to beginning the project, the team of Drs. Kitchener, King, Wood, and Lynch met to design the computerized assessment instrument for Reflective Judgment. During the summer and fall, Dr. Wood and a graduate student designed two modular

computer programs that could ideally be used as a framework for the computerized assessment. The programs took into consideration several theoretical and psychometric considerations. For example, stage prototypic examples were written and presented in a window format that allowed participants to move back to earlier parts of the measure, the program was designed so that participants had to scroll through all the options before choosing one to counteract the temptation to choose the first attractive option, and response times were monitored. The last option allowed data from participants with response times that appeared particularly short (and may have reflected inattention to the statements in the measure) to be separated for analyses of the data. Effort also was made to generate a computer program that was portable to a variety of computer configurations. A software program also was developed to analyze the raw data from the computerized assessment. During the first year the computerized measure was piloted once, revised after a meeting of the team at BGSU, piloted and revised again, and finally, repiloted on a sample of University of Missouri freshmen and seniors (with the financial support of the UM-C Psychology Department).

At the end of the first year, we concluded that the new discipline-specific problems for chemistry and business performed well in the interview format. Students scored very similarly on the new discipline-specific problems as they did on the traditional RJI problems. However, we were unable to get the reliability of the computerized measure above .50. We concluded that moving directly from an interview format to an objective, computerized format for assessing Reflective Judgment was too great a leap to accomplish in a single step. After the many revisions to the computerized format and further pilot testing, statistical analyses of the results indicated repeatedly that the psychometric properties of the instrument were too poor to provide reliable information about the reflective thinking of the students who responded to the instrument.

Our investigations indicated that the computerized assessment strategies did not provide students with enough structure and support to yield data that could be evaluated reliably. The interview format allows persons to first state their point of view about a particular problem and then, with careful and individually appropriate probing by the interviewer, to talk about their

underlying assumptions about knowledge and justification of their beliefs more specifically. The initial attempt to computerize the assessment did not allow those responding to employ a step-by-step approach to thinking about the questions at hand. In addition, because all response options could not appear on the screen at the same time, it appeared that students had difficulty evaluating different choices when they had to scroll back and forth between screens to read their options. We also discovered that although some educators believed the computerized format would be exciting and utilitarian, many others would prefer a paper and pencil format that would fit better with the other types of assessment commonly used by postsecondary institutions.

Because of the difficulties associated with the computerized format and the expense of computer programming, the project team stepped away from pursuing the development of that format during the rest of the project period. Instead, the team focused its instrument development resources on devising a paper and pencil form that more closely mirrored the interview format. This shift occurred in Year 2 of our 3 year project, and the instrument was named the Reflective Thinking Appraisal (RTA, see Appendix A).

In its current form, the RTA includes two problems from the Reflective Judgment Interview: one that addresses historical knowledge, specifically how the pyramids were built, and another that addresses scientific knowledge, the safety of artificial sweeteners. Most students can complete the instrument in about an hour, although many take only 30 to 40 minutes to complete it. Persons are asked to read a problem and select a point of view that is closest to their own. Then they are guided through a set of tasks related to each of four questions drawn from the Reflective Judgment Interview. These questions center around (1) the basis or explanation for one's point of view, (2) how sure one is about the correctness of that point of view, (3) whether or not one point of view is correct and another is incorrect when there are disagreements about the problem, and (4) how it is possible for two experts in the field who have the same information to genuinely disagree on this issue and arrive at different conclusions. For each set of questions, persons are asked to write a brief statement in answer to the question. The task then is to compare what they were thinking when they wrote that response to a series of statements that

reflect qualitatively different explanations that can be coded by stage based on the Reflective Judgment model. Respondents are asked to mark each statement in one of four categories: very similar, similar, dissimilar, or very dissimilar. Finally, persons are asked to rank the three statements that are most similar to their own written response to the question. It is these rankings that are used to score the RTA.

During Year 2 the RTA was initially pre-piloted on a small sample ($n$=18). After judging that it performed well and revising it slightly, it was again piloted on a sample of over 350 students from seven sites. The sample included high school, undergraduate, and graduate students. Analyses indicated that reliability across the educational age range tested was .61 and that scores for older, more educated students were significantly higher than for younger, less educated students.

No funds were budgeted for additional piloting of the measure during Year 3 (and an extension into Year 4), but because the reliability of the measure was still below generally accepted standards, the decision was made to use savings from the Year 2 budget and support from the University of Missouri to again revise and pilot the RTA. At this time, the measure was piloted at the three primary project institutions on a sample of 534 students. Our current data indicate that differences between younger, less educated students and older, more educated students are significant, and the reliability over the entire age range is estimated at .79. (See the technical manual in Appendix B for a complete description of the psychometric properties of the RTA.)

**Goals 3 and 4: Faculty Consultation.** At the end of the first year of the project, we invited the consultants identified earlier in this report to join us for a retreat. These persons were selected because they had already been using the Reflective Judgment model as a tool in designing the courses they teach. The focus of this retreat was to learn from these experienced faculty about how they used the developmental model in their teaching, what difficulties and concerns they had encountered in their efforts, and what advice they had for us and other faculty who

pursue similar endeavors. From what we learned at the retreat, we designed 3-day workshops for faculty participants at DU and BGSU.

Our project-related work with faculty on our own campuses was initiated when we asked interested business faculty at DU and chemistry faculty at BGSU to participate in the development of ill-structured problem statements for use in the Reflective Judgment Interview format. With these core groups as a starting point, larger groups of interested faculty were identified and asked to participate in the "Reflective Judgment in the Classroom" workshops. Drs. Kitchener and Lynch and our consultant, Dr. Nevins, conducted a workshop with 8 faculty from DU's College of Business Administration. Dr. King and our consultant, Dr. Finster, worked with 10 faculty from a variety of disciplines at BGSU. As the workshops progressed, we attempted to adapt the content and format to the needs of the participants. Daily evaluation forms were one source of information about areas of success and participant needs (see Appendix C for an example). The workshops were generally successful, and we continued to meet and consult with the participating faculty, both in groups and individually, over the next year as they designed course revisions and implemented new teaching strategies.

In addition to these two workshops with small groups, the project team and consultants made several presentations at the previously mentioned UM-C Wakonse retreat. Each participant in the retreat wrote goals for him/herself at the end of the retreat, and 42 of those goals were related to Reflective Judgment. Dr. Wood also spoke about using the Reflective Judgment model in the classroom to 45 persons from UM-C who attended the two-day Annual Teaching Renewal Conference. He made a presentation about a similar topic to over 300 persons at the University of Tennessee at Martin.

## RESULTS

**Goals 1 and 2: Assessment Instrument.** Our latest data indicate that the written assessment instrument, the RTA (see Appendix A), is performing fairly well. Older, more educated groups score higher than younger, less educated groups, and the overall reliability is

approaching .80. See Tables 1 and 2 for data about the norming sample (*n*'s and mean age of subsamples), Table 3 for coefficient alpha internal consistency estimates, and Figure 1 for a visual representation of RTA scores of different education level subsamples. Dr. Wood has written a scoring program for the instrument and has produced the *Technical Manual to Accompany the Reflective Thinking Appraisal* (see Appendix B). The technical manual describes in detail the uses and performance of the RTA. Although at this point in time only two dilemmas from the standard Reflective Judgment Interview have been converted to the paper and pencil format, other discipline-specific problems in business, science, and psychology have been successfully piloted in the interview format. These problems (stated in King & Kitchener, 1994) stand ready for use as we continue to develop the RTA. More will be said about our future plans for instrument development in the section entitled "Next Steps."

**Goals 3 and 4: Faculty Consultation**.

**University of Denver.** Questionnaires and interviews with participating faculty from the DU's College of Business Administration revealed that seven out of eight identified attending to the needs of their students for support in developing critical thinking skills was the most important thing they learned during the project. For example, they suggested that understanding the developmental process of learning to make reflective judgments helped them to more appropriately structure and sequence course activities. Better structure and sequence of course activities provided both intellectual and emotional support for students who were being asked to address difficult course-related problems. Several faculty reported that attending to support issues improved their relationships with students and thus increased student's participation and interest in the FIPSE target courses. Providing more deliberate support for students' efforts was viewed as an important concept that they believed would continue to have a positive impact on their teaching.

More than a year after the initial workshop, this group of faculty reported that they continued to talk among themselves and with colleagues who had not participated in the FIPSE project about ill-structured problems in their disciplines and how to help students address complex

problems in their courses and across the College of Business Administration. FIPSE project personnel were asked to conduct a Reflective Judgment workshop for the entire School of Accountancy faculty and to present a program to the Colorado Society of Certified Public Accountants Faculty/Practitioner Roundtable. Information and experiences gained during the FIPSE project also were employed as the College of Business Administration revamped the MBA curriculum into an "integrated" program with a focus on problem solving in the business world.

A substantial body of empirical data previously gathered using the Reflective Judgment Interview on DU business students provided faculty with information about the range of Reflective Judgment stages they might expect their students to exhibit. Although we did not expect significant increases in students' Reflective Judgment scores after a single, quarter long target course, we had originally hoped to have discipline-specific, objective assessment instruments available for use during the faculty consultation portion of this project. Data obtained at the beginning of the target courses with such instruments could have provided faculty with helpful information about the developmental characteristics of their students. Because the instrument development process was slower than we anticipated, we instead encouraged faculty to develop their own, course-appropriate informal Reflective Judgment assessment assignments (for an example, see Appendix D, *Developing Reflective Judgment in the Classroom: A Manual for Faculty*). In classes in which this strategy was used, the student essays yielded information that could be roughly divided into qualitatively different groups by an experienced, certified Reflective Judgment rater based on the Reflective Judgment Scoring rules. These data were then used to help faculty members better understand the developmental needs of their students. At this point in time, we have no empirical data to indicate whether or not students involved in courses adapted based on the Reflective Judgment model and the concepts of appropriate challenge and support improve significantly in Reflective Judgment scores. See the "Next Steps" section for information about current efforts to assess this type of outcome.

Although most DU faculty who participated in the FIPSE project reported increased student participation and interest in the target courses, this was not empirically investigated and

substantiated. In fact, the results of students' evaluations of the target courses via the standard College of Business Administration assessment instrument were mixed. Some of the teachers and courses received improved ratings, others remained about the same as earlier assessments, but at least one received poorer ratings. Initially poorer ratings are a significant risk associated with innovation. In the process of making changes, some changes will be helpful and productive, and others will not be as effective. It also is likely that those standard course evaluation forms do not adequately tap what students have learned, but reflect how much students liked the course or how they perceived the comfort level of the professor regarding the course.

**Bowling Green State University.** At BGSU, at least one class period of each of the 10 faculty participants was observed by the FIPSE project staff. These observations, as well as the written reports of the faculty participants, revealed that faculty were using a variety of strategies to encourage their students to think more reflectively. The strategies included informal assessment techniques such as journal writing, providing opportunities for students to examine ill-structured problems from multiple perspectives, and providing support for students to justify their positions more fully.

Faculty explicitly stated that they were listening to their students more intently and differently than they had before the faculty workshop. They had learned to use the Reflective Judgment model as a tool for understanding student responses to ill-structured problems. The BGSU faculty participants reported that the project will have an on-going impact on their teaching. One participant wrote, "My general attitude toward teaching has also changed. In the future, it will be hard for me to teach any course--particularly on the undergraduate level--without considering both the level of reflective thinking at which my students are likely to be at the outset of the course and the level which I'd like them to attain at the end."

The interdisciplinary nature of the group made it different from the group of business faculty at DU, but at least one BGSU participant considered this a significantly positive factor: "It was refreshing to gather a variety of faculty from different disciplines to engage seriously in discussion of cognitive growth and teaching. This rarely happens." Dr. King observed that the

individuals in this faculty group brought a wealth of experience and an interesting mix of perspectives to the group discussions, which set the stage for a lively and productive exchange of ideas.

The faculty participants at BGSU gathered information from their students about their perceptions of the revised courses. In a science course, students were challenged to move away from the pervasive view held by most of them that science is dominated by ideas about "observable facts and proven theories." The professor reported that "at the beginning of the course, only two of the 26 students demonstrated an appreciation of the tentative nature of scientific understanding. Some changes were apparent by the end of the course." Students in other courses were challenged to examine and support their own beliefs, and to consider alternative points of view as potentially valid. They reported becoming more aware of their own thought processes. "I learned to view issues with an open mind and to consider them before forming opinions," observed one student. According to another, [I am] "less likely to accept a thought or idea without reasoning it out myself and being able to support it." Statements like these indicate that students were being confronted with the idea of thinking more reflectively and were making progress toward doing so.

**Summary**. The FIPSE project team learned several significant things about consulting with faculty as a result of our work on this project. Faculty are like university students in that they, too, need appropriate challenges and support as they work on the ill-structured problem of improving their teaching. Faculty, like students, come to a consultative experience with varying agendas, interests, and skills, and some grasp what a consultant is trying to convey rather quickly, other less completely and more slowly.

This experience has helped us fine tune our presentations to faculty for workshop formats and presentations at professional meetings. For example, these changes include adding more examples specific to target disciplines, providing more hands-on experience and individual feedback for faculty when time permits, and making clearer at the outset what expectations are appropriate regarding students' changes in reflective thinking. We also emphasize the need for

faculty to devise informal Reflective Judgment assessment activities that are specific to their course content, and we explicitly address the potential discomfort of both faculty and students about these subjective assessment strategies.

The FIPSE project team also learned that faculty development efforts designed to do more than just initiate interest require much more than a 3-day workshop. After the workshops, we scheduled and conducted follow-up meetings with our FIPSE faculty groups and consulted individually with faculty as time permitted. We also observed some of the faculty in the classroom; others did not invite us into their classrooms. Future projects with a long-term faculty consultation component might include the expectation that consultants will observe in the classroom as part of the initial agreement for participation and that the results of department course evaluation forms will be made available to the consultants. This would greatly strengthen the evaluation component of such projects.

Departmental and institutional support for faculty development efforts also are needed for successful consultation. Faculty often commented that they rarely have time to talk in depth about teaching issues with their colleagues. Some reported that taking time to try to improve their teaching was perceived by other faculty members as inappropriate, given the competing demands for research and publications and the structure of the "faculty reward system" which focuses on non-teaching activities. For some faculty, the only information about their teaching that is considered in promotion and tenure deliberations is gleaned from the empirical analysis of students' course evaluations. As noted earlier, a narrowly focused approach to evaluating teaching may actually work against faculty who are experimenting with innovations in their courses.

**NEXT STEPS**

The work funded by this FIPSE project will continue beyond the end of our grant period. Development of the RTA assessment instrument will proceed as financial resources and data sources are available. In addition, on-going faculty consultation will allow us to use the RTA to

assess the long-term effects of developmental instructional strategies on students' thinking. For example, at the University of Denver, we have used the information learned in our earlier faculty consultation efforts to work with six faculty and five teaching assistants. These faculty and teaching assistants, with the support of a National Science Foundation (NSF) grant, have designed and are teaching core science classes in a way that is intended to promote critical thinking about ill-structured problems in science. The RTA along with the Reflective Judgment Interview are being used in the 1993-94 academic year to pretest and posttest students who have taken a 3-quarter sequence of science core courses for nonscience majors. It will provide valuable information about student development and the impact of efforts specifically designed to foster reflective thinking. This investigation also includes a quasi-experimental design component that will allow comparisons of students in the "target" reflective judgment courses with those in courses not specifically designed to promote reflective judgment. In addition, these data will allow us to further refine the RTA by evaluating the correlation between RTA scores and Reflective Judgment Interview scores, as well as with SAT/ACT scores. In essence, the work we are doing with the NSF grant is allowing us to combine the faculty consultation and student assessment work that we originally envisioned occurring in the second year of our FIPSE grant.

We also are consulting with Dwight Giles at Vanderbilt University about the evaluation component of his FIPSE grant that is investigating service learning outcomes. Pilot pre- and posttests using the RTA were conducted this spring in conjunction with that FIPSE project. Further, as noted earlier, our work continues with the College of Business Administration at the University of Denver to integrate reflective judgment into the curriculum.

When we are confident that the format of the RTA is viable, we hope to add problems from other disciplines to the RTA problems set. Our FIPSE project funded the successful interview piloting of problems in chemistry, business, and psychology. If funds become available, we also hope at some point to reconsider the option of developing a computerized version of the RTA as well.

Our dissemination efforts also will continue. For example, Dr. Lynch is scheduled to make presentations at the American Association of Higher Education's (AAHE) Conference on Assessment and Quality in June, 1994. We will continue to respond to all inquiries about our objective measure of Reflective Judgment and to consult with faculty who are interested in helping their students think more reflectively. As the network of interested educators grows, we hope to conduct more workshops and consult with educators, researchers, and administrators on other campuses.

## SUMMARY AND CONCLUSIONS

Four purposes were addressed in this three year FIPSE project:

1. To develop an objective measure of Reflective Judgment that could be used by faculty and administrators to evaluate the nature of students' reasoning about ill-structured problems.

2. To adapt the Reflective Judgment Interview to the specific content of three disciplines: business, chemistry, and psychology.

3. To use data on Reflective Judgment already collected to consult with faculty about how to adapt their instruction to the developmental characteristics of students.

4. To develop materials that would allow faculty at other sites to use the objective measure and to apply the Reflective Judgment model to their teaching.

The goals of the project were somewhat different from those originally proposed, but the focus of the project on developing an objective instrument to assess Reflective Judgment and on helping faculty more deliberately encourage reflective thinking in the classroom remained paramount.

The project resulted in a viable version of the Reflective Thinking Appraisal, a paper and pencil instrument designed to assess how students think about ill-structured problems. A technical manual to accompany this assessment instrument also is available. The project also

helped small groups of faculty at the University of Denver, Bowling Green State University, and the University of Missouri--Columbia to redesign courses so that students are deliberately encouraged to improve their reflective thinking skills. A manual for use by faculty was written based on the project team's consultation with participating faculty.

In addition to the assessment instrument, the technical manual, and the faculty manual, the project team gained two generally valuable insights: First, developing sound, well-validated instruments to assess critical thinking is a long-term project. Even though those efforts for this project were founded on a valid theoretical model and a much used interview assessment strategy, three and a half years was enough time for only the initial stages of the instrument development. Second, efforts at truly successful faculty development require sustained interest and work on the part of both the participating faculty and the consultants. A three-day workshop can only put the process of curricular change in motion. Continued contact among faculty groups and between faculty and consultants is an important part of the long-term work of curricular adaptation to encourage reflective thinking.

The project set the stage for continuing the development of the RTA and expanding its use to other campuses and projects. It also laid the foundation for expanding the faculty consultation to other institutions which currently is occurring. The faculty manual can be used individually by interested educators, but it is specifically designed to be used in conjunction with workshops or other presentation formats conducted by members of the project team. The ultimate success of this project can be assessed only in the months and years to come as we continue the instrument development and faculty consultation efforts initiated with this grant.

Table 1.

Sample Sizes for Norming Sample by Site and Educational Level

| Educational Level | N | Inst. 1 (BGSU) | Inst. 2 (DU) | Inst. 3 (MO) | Other -- |
|---|---|---|---|---|---|
| Undergraduate | | | | | |
| Freshmen | 188 | 23 | 3 | 162 | |
| Sophomore | 65 | 26 | 2 | 37 | |
| Junior | 40 | 18 | 6 | 16 | |
| Senior | 114 | 14 | 7 | 93 | |
| Graduate | | | | | |
| Currently completing: | | | | | |
| 1st half of master's course work | 20 | | 4 | 16 | |
| 2nd half of master's course work | 21 | 0 | 1 | 20 | 2 |
| 1st half doctoral course work | 21 | 1 | 3 | 15 | 2 |
| 2nd half doctoral course work | 36 | 5 | 10 | 19 | 2 |
| Doctoral Coursework Completed | 29 | 4 | 6 | 17 | |
| Total | 534 | 91 | 42 | 395 | |

RTA Technical Manual June 6, 1994 2:02 pm Page 8
©1994 Reflective Thinking Associates

Table 2.

Age Means for Norming Sample by Site and Educational Level

| Educational Level | Overall | Inst. 1 (BGSU) | Inst. 2 (DU) | Inst. 3 (MO) | Other -- |
|---|---|---|---|---|---|
| Undergraduate | | | | | |
| Freshmen | 18.5 | 18.3 | 20.7 | 18.5 | |
| Sophomore | 19.8 | 19.9 | 21.5 | 19.7 | |
| Junior | 20.2 | 20.2 | 20.3 | 20.2 | |
| Senior | 22.8 | 23.6 | 25.9 | 22.4 | |
| Graduate | | | | | |
| Currently completing: | | | | | |
| 1st half of master's course work | 26.7 | - | 27.8 | 26.4 | |
| 2nd half of master's course work | 28.5 | - | 26.0 | 28.7 | 38.5 |
| 1st half doctoral course work | 31.3 | 39.0 | 38.7 | 28.3 | 31.0 |
| 2nd half doctoral course work | 32.0 | 27.5 | 36 | 31.9 | 38.0 |
| Doctoral Coursework Completed | 33.1 | 31.8 | 37.3 | 31.3 | |
| Total | 22.7 | 29.6 | 22.1 | 21.2 | |

RTA Technical Manual June 6, 1994 2:02 pm Page 10
©1994 Reflective Thinking Associates

Table 3

Coefficient Alpha Internal Consistency Estimates by Educational Level & Site[2]

| Educational Level | Overall | Inst. 1 (BGSU) | Inst. 2 (DU) | Inst. 3 (MO) | Other -- |
|---|---|---|---|---|---|
| Undergraduate | | | | | |
| Freshmen | .62 (.68) | .88 (.87) | | .48 (.59) | |
| Sophomore | .68 (.76) | .58 (.62) | | .70 (.80) | |
| Junior | .63 (.61) | .83 (.48) | | .35 (.83) | |
| Senior | .71 (.70) | .83 (.82) | | .67 (.63) | |
| Total Undergraduate | .67 (.70) | .76 (.75) | | .60 (.67) | |
| Graduate | | | | | |
| Currently completing: | | | | | |
| 1st half of master's course work | .83 (.85) | | | .79 (.83) | |
| 2nd half of master's course work | .73 (.75) | | | .78 (.75) | |
| 1st half doctoral course work | .76 (.71) | | | .84 (.77) | |
| 2nd half doctoral course work | .64 (.55) | | | .57 (.47) | |
| Doctoral Coursework Completed | .82 (.83) | | | .86 (.90) | |
| Total Graduate[3] | .76 (.74) | | | .78 (.75) | |
| Grand Total | .77 (.79) | | | | |

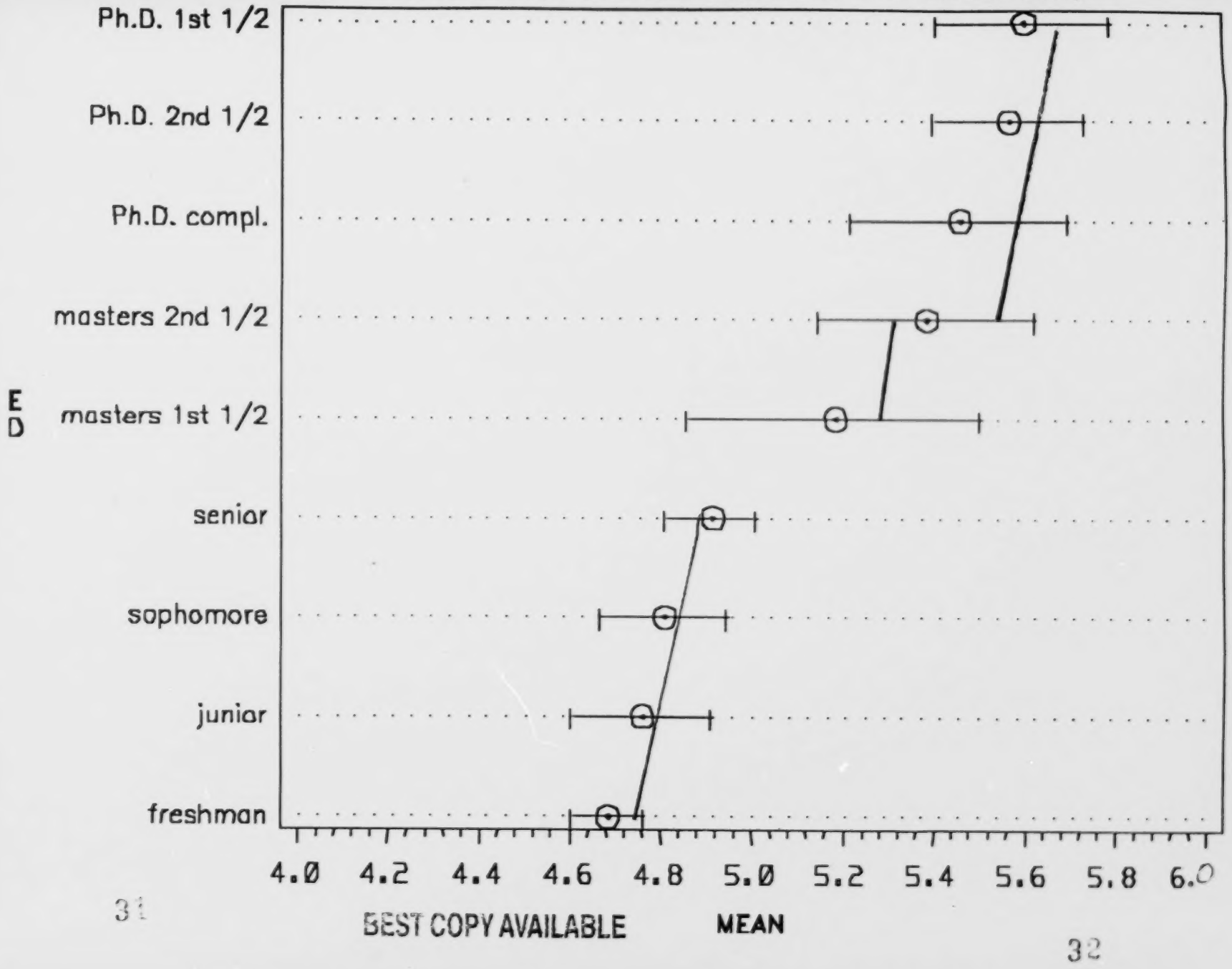
2. Numbers in parentheses indicate estimates derived when individuals with high rates of meaningless statement endorsement are excluded

3. Excluding Individuals with Doctoral Coursework Completed


©1994 Reflective Thinking Associates

BEST COPY AVAILABLE

Figure 1: Overall RTA score by Educational Level (Based on ind.'s w/mean.<.05)






BEST COPY AVAILABLE

APPENDIX A

DO NOT USE OR DUPLICATE
DEVELOPMENT DRAFT

# Reflective Thinking Appraisal

## Introduction

This questionnaire is aimed at understanding how people like you think about various issues; it asks not only **what** you think, but **why** you hold the opinions you do. People give different responses to the questions asked here, so please give the best answer **you** have to each question.

NOTE: If a statement does not make sense to you (and some probably will not make sense), mark this statement as "very dissimilar." This is to make sure you are reading the statements carefully.

**Take as long as you need to read these instructions and answer the questions.**

Each section of this questionnaire begins with a description of a particular issue, followed by several questions about that issue. Please answer each question in order, proceeding through the questionnaire step-by-step from the first through the last page.

**Please circle one:** Male Female

**Date of Birth:** _____ **Name of School/College:** _____

**Date of Testing:** _____ **Academic Major (if known):** _____

**Current Student Status (please check one):**

| High School | Undergraduate | Graduate |
|---|---|---|
| ____ 9th grade | ____ Freshman | ____ 1st half master's course work |
| ____ 10th grade | ____ Sophomore | ____ 2nd half master's course work |
| ____ 11th grade | ____ Junior | ____ 1st half doctoral course work |
| ____ 12th grade | ____ Senior | ____ 2nd half doctoral course work |
| | | ____ Doctoral course work completed |

[Code #_____]

©Reflective Judgment Associates, 1993
Version C | Revised: September 7, 1993



33

# Issue A

**Please note the current time:** _____

People often have to make decisions that may affect their health. One example is deciding whether to eat something containing chemical additives, because there have been conflicting reports about the relationship between chemicals that are added to foods and the safety of these foods. For example, some studies have indicated that even in small amounts, artificial sweeteners (such as Nutrasweet™) can cause health problems, making foods containing them unsafe to eat. Other studies, however, have indicated that even in large amounts, artificial sweeteners do not cause health problems, and that the foods containing them are safe to eat.

1. **Which of the following responses is closest to <u>your</u> point of view about this issue?**

   ❑ I think that foods containing artificial sweeteners are safe for people to eat.

   ❑ I do ***not*** think that foods containing artificial sweeteners are safe for people to eat.

**Instructions for Question 2**

***THINK*** about your own response to this question, then ***WRITE DOWN*** your answer using the space below.

**2. People give different explanations for their point of view about the safety of artificial sweeteners in foods. What is the basis for <u>your</u> point of view about this question?**

_____

_____

_____

_____

**Instructions for Question 3**

Next is a list of other possible responses to the first question. ***READ*** each statement carefully. ***DECIDE*** whether it is "very similar," "similar," "dissimilar," or "very dissimilar" <u>to the approach you used to answer this question</u>. **Place an "X" in the appropriate box next to each statement.** If none of the statements <u>exactly</u> resembles your response, choose the one that <u>best</u> reflects your approach to answering this question. If a statement does not make sense to you (and some probably will *not* make sense), mark this statement as "very dissimilar." (This is to make sure you are reading the statements carefully.)

**3. How similar is each of the following statements to the <u>approach you used</u> to answer Question 2?**

| VS | S | D | VD | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | Ⓐ There isn't much proof on either side of the issue about the safety of artificial sweeteners in foods, so I believe what I want to believe. My point of view just makes sense. |
| ☐ | ☐ | ☐ | ☐ | Ⓑ After comparing the interpretations on both sides of the issue, my point of view seems more reasonable to me because the evidence is stronger and the assumptions on which this view is based seem more valid. |
| ☐ | ☐ | ☐ | ☐ | Ⓒ When I hear a scientist say whether an artificial sweetener is safe or not, then I know what to believe. |
| ☐ | ☐ | ☐ | ☐ | Ⓓ I look at the quality and density of the proof-claim of this issue and base my assumptions accordingly. The facts of this issue must be probabilistically migrated from that which is unproven to proven. |
| ☐ | ☐ | ☐ | ☐ | Ⓔ My point of view is based on my analysis of where the weight of the evidence lies. It is more probable because it best accounts for the evidence and other things I know about related topics, such as nutrition. |
| ☐ | ☐ | ☐ | ☐ | Ⓕ The facts aren't very clear because there are so many variables involved in assessing the safety of artificial sweeteners in foods. So I just believe what seems right to me about their safety. |
| ☐ | ☐ | ☐ | ☐ | Ⓖ There are several valid ways of looking at this issue. People's conclusions are related to their assumptions about how scientists do research and draw conclusions; people's assumptions determine how they interpret evidence. |

**Instructions for Question 4**

*CHOOSE* the **one** statement from those listed on p. 4 that most closely resembles the approach you use in thinking about this question. *WRITE* the **letter** of that statement (A, B, C, etc.) in the circle below the question. Again, if none of the statements exactly resembles your response, choose the one that best reflects your approach to answering this question.

Then *CHOOSE* the statement from p. 4 that is the next most similar to the approach you use and *WRITE* that letter in the second circle.

Then *CHOOSE* the statement from p. 4 that is third most similar to the approach you use and *WRITE* that letter in the third circle.

NOTE: The statements you choose as being *more like how you think* should be checked in a column further to the left than the statements ranked as being *less like how you think*.

For example, if you marked two statements in the "very similar" column, you would re-read each of these and indicate in the first circle which one is **most** like how you think. Then rank the other as **second most** like how you think. Next, you would look at those statements marked as "similar" to determine which statements is **third most** like your own thinking about this question.

If you marked only one statement as "very similar," list that statement in the first circle as most like what you think. Then look at the statements marked as "similar" for the second ranking, and so on.

4. **Which three statements (A, B, C, etc.) from the list above are most similar to your approach? Please indicate these in rank order below.**

Statement ◯ is most like how I think.

Statement ◯ is second most like how I think.

Statement ◯ is third most like how I think.

**NOTE: Use the same instructions given on pages 4 & 5 for the remaining questions.**

**Please turn the page and continue, answering the questions thoughtfully and in the order in which they are presented.**

5. **Some people think they already know for sure about the safety of artificial sweeteners in foods. Other people don't know for sure about this. What do you think? Why?** *(Write your answer in the space below.)*

6. ❑ I *don't* know for sure about this. ***[If you mark this answer, continue to Question 9, p. 7.]***

❑ I know for sure about this (or am fairly sure). ***[If you mark this answer, continue to Question 7, below.]***

7. **How similar is each of the following statements to the approach you used to answer Question 5?**

| VS S D VD | | **VS**=Very Similar; **S**=Similar; **D**=Dissimilar; **VD**=Very Dissimilar |
|---|---|---|
| ☐☐☐☐ | Ⓐ | I am sure about the safety of foods containing artificial sweeteners based on what I have been taught. |
| ☐☐☐☐ | Ⓑ | Knowing for sure is a pretty relative thing. It depends on your criteria for accepting an interpretation as certain. |
| ☐☐☐☐ | Ⓒ | I am sure about the safety of foods containing artificial sweeteners because it just feels right to me and I just believe what I want to about this matter. |
| ☐☐☐☐ | Ⓓ | I am sure because I have evaluated the evidence and its fit with related arguments and assumptions. As a result of that evaluation, I am confident about the validity of my conclusion. |
| ☐☐☐☐ | Ⓔ | I am sure because my confidence impacts my single, infinite assessment of the underlying assumptions of this issue prior to collaborating data. |
| ☐☐☐☐ | Ⓕ | I am fairly certain because I have constructed a reasonable interpretation of the evidence, but I may not have interpreted the evidence adequately. |
| ☐☐☐☐ | Ⓖ | I am sure for myself based on the facts I know, but other people may be just as sure of a different opinion based on the facts they know, and they have a right to their own opinion about this question. |

8. **Which three statements (A, B, C, etc.) from the list above are most similar to your approach? Please indicate these in rank order below.**

Statement ◯ is most like how I think.

Statement ◯ is second most like how I think.

Statement ◯ is third most like how I think.

**If you completed Questions 7 & 8,**
**GO TO QUESTION 11 ON PAGE 8**

**9. How similar is each of the following statements to <u>the approach you used</u> to answer Question 5 (p. 6)?**

| VS | S | D | VD | | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | Ⓐ | I don't know for sure because there are many variables involved and there is no way to research their effects on everyone. I do, however, have a personal opinion about this based on what I believe about the evidence. |
| ☐ | ☐ | ☐ | ☐ | Ⓑ | I don't know for sure because I am not confident and because of the superlative opinions I hold about this point of view and the discrepant assumptions I can draw from this view. |
| ☐ | ☐ | ☐ | ☐ | Ⓒ | I don't know for sure about the safety of foods containing artificial sweeteners, but I could easily find out by simply asking someone who does know. |
| ☐ | ☐ | ☐ | ☐ | Ⓓ | I don't know for sure right now about the safety of artificial sweeteners in foods because experts like scientists don't yet have all of the information. But when they do more work, they will know for sure whether or not they are safe. |
| ☐ | ☐ | ☐ | ☐ | Ⓔ | I don't know for sure because there are several valid interpretations of the evidence. However, after comparing these interpretations, I could construct a reasonable point of view based on its consistency with the available evidence. |
| ☐ | ☐ | ☐ | ☐ | Ⓕ | I don't know for sure because people's interpretations of the evidence are always affected by their background, training and assumptions. Thus, we have different ways of evaluating the evidence. |
| ☐ | ☐ | ☐ | ☐ | Ⓖ | I don't know for sure, but I think we can come very close to being sure. Even though there are competing interpretations of the evidence, I can conclude that one view is more probable because it best accounts for the evidence or because its underlying assumptions are more plausible. |

**10. Which three statements (A, B, C, etc.) from the list above are <u>most similar</u> to your approach? Please indicate these in rank order below.**

Statement ◯ is <u>most</u> like how I think.

Statement ◯ is <u>second</u> most like how I think.

Statement ◯ is <u>third</u> most like how I think.

BEST COPY AVAILABLE

**11. When two people disagree about the safety of artificial sweeteners in foods people eat, is one person's point of view correct and other person's view incorrect? Why? Why not?** *(Write your answer in the space below.)*

_____

_____

_____

_____

**12.** ❑ No, one view is *not* correct (or mostly correct). ***[If you mark this answer, <u>continue to Question 15, p. 9.</u>]*** ⇨

⇩ ❑ Yes, one view is correct (or mostly correct). ***[If you mark this answer, <u>continue to Question 13</u>, below.]***

**13. How similar is each of the following statements to <u>the approach you used</u> to answer Question 11?**

| VS S D VD | **VS**=Very Similar; **S**=Similar; **D**=Dissimilar; **VD**=Very Dissimilar |
|---|---|
| ☐☐☐☐ Ⓐ | Yes, that person's point of view would be correct for that person. But for someone else, a different point of view would be correct. |
| ☐☐☐☐ Ⓑ | Yes. One person would really know which point of view is correct about the safety of artificial sweeteners in foods; the other would just be wrong. |
| ☐☐☐☐ Ⓒ | Yes. By correct, I mean that some explanations are very far along the continuum of the probable effects of artificial sweeteners. It is a matter of choosing the position that seems most correct, most in line with the evidence, the best explanation. |
| ☐☐☐☐ Ⓓ | Yes. One person's view is probably correct, but we just don't know which one of these is correct right now. Some day I'm sure we'll know which one is correct. |
| ☐☐☐☐ Ⓔ | Yes, there is a correct answer, but it may not be knowable. We can only compare the interpretations and decide which seems more reasonable or accurate. |
| ☐☐☐☐ Ⓕ | Yes. You could say that one point of view is correct, but that evaluation would be relative to a particular way of understanding the issues surrounding the safety of artificial sweeteners in foods. |
| ☐☐☐☐ Ⓖ | Yes, because the rule for spontaneous consensual criticism offers a premeditated basis for choosing whether artificial sweeteners are safe. |
| ☐☐☐☐ Ⓗ | Yes, but not in an absolute sense. One point of view may be more sensible or more scientifically accurate than the other, but we can never know whether it is absolutely correct or not. |

**14. Which three statements (A, B, C, etc.) from the list above are <u>most similar</u> to your approach? Please indicate these in rank order below.**

Statement ◯ is <u>most</u> like how I think.

Statement ◯ is <u>second</u> most like how I think.

Statement ◯ is <u>third</u> most like how I think.

BEST COPY AVAILABLE

**If you completed Questions 13 & 14,**
**GO TO QUESTION 17 ON PAGE 10**

**15. How similar is each of the following statements to <u>the approach you used</u> to answer Question 11 (p. 8)?**

| VS | S | D | VD | | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | Ⓐ | No, you cannot say one point of view is correct because a person's evaluation is relative to a particular way of understanding the problem. |
| ☐ | ☐ | ☐ | ☐ | Ⓑ | No, not in an absolute sense. One point of view may be more feasible or more scientifically accurate than the other, but we can never know absolutely whether or not it is correct. |
| ☐ | ☐ | ☐ | ☐ | Ⓒ | No, not until one is proven. Until it is proven, it is just a matter of what you want to believe about whether artificial sweeteners are safe or not. |
| ☐ | ☐ | ☐ | ☐ | Ⓓ | No, you can't say one is correct and one is incorrect because there is no definite proof; they're both just opinions about the facts. What each person thinks is an individual thing. |
| ☐ | ☐ | ☐ | ☐ | Ⓔ | No, because there isn't that much proof about whether artificial sweeteners are safe or not, so either view could be right or wrong. Until they prove it, you can't say which is correct. |
| ☐ | ☐ | ☐ | ☐ | Ⓕ | No, because we can't say which point of view is correct. It is so hard to prove that something is definitely correct because we don't understand all of the variables involved. |
| ☐ | ☐ | ☐ | ☐ | Ⓖ | No, because we can never have all the evidence or know if the information is being interpreted correctly. Nevertheless, we can judge one point of view as more plausible given what we now know. |
| ☐ | ☐ | ☐ | ☐ | Ⓗ | No, I cannot say absolutely that one view is correct because the evidence is so complex and open to interpretation, but I can evaluate one point of view as being more correct than another based on the evidence I have. |
| ☐ | ☐ | ☐ | ☐ | Ⓘ | No, because being correct requires interpreting how well informed your facts and other criteria are for being knowledgeable about the safety of artificial sweeteners. |

**16. Which three statements (A, B, C, etc.) from the list above are <u>most similar</u> to your approach? Please indicate these in rank order below.**

Statement ◯ is <u>most</u> like how I think.

Statement ◯ is <u>second</u> most like how I think.

Statement ◯ is <u>third</u> most like how I think.

BEST COPY AVAILABLE

17. **How is it possible that two experts in the field who have the same information can genuinely disagree on this issue and arrive at different conclusions?** *(Write your answer in the space below.)*

18. **How similar is each of the following statements to the approach you used to answer Question 17?**

| VS | S | D | VD | | |
|---|---|---|---|---|---|
| **VS**=Very Similar; **S**=Similar; **D**=Dissimilar; **VD**=Very Dissimilar | | | | | |
| ☐ | ☐ | ☐ | ☐ | Ⓐ | Real experts who are honest will not disagree. |
| ☐ | ☐ | ☐ | ☐ | Ⓑ | Experts disagree because they approach the problem with different conclusions already in mind and then find evidence to support their conclusions. |
| ☐ | ☐ | ☐ | ☐ | Ⓒ | Experts disagree because their beliefs are relative to their own perspective. As a result, they interpret the evidence differently. |
| ☐ | ☐ | ☐ | ☐ | Ⓓ | Experts disagree because it is not yet known whether artificial sweeteners are safe. Until there is more evidence, they will believe whatever they want to believe about it. |
| ☐ | ☐ | ☐ | ☐ | Ⓔ | Experts disagree because the premeditated hard evidence is synthesized into available belief systems about different comprehensive factual analyses. |
| ☐ | ☐ | ☐ | ☐ | Ⓕ | Experts disagree because they begin with different beliefs and experiences. These lead them to look at the facts differently. |
| ☐ | ☐ | ☐ | ☐ | Ⓖ | Experts disagree as a result of differences in assumptions, emphasis, and methods of evaluating and interpreting evidence. However, the adequacy of their conclusions can also be evaluated. |
| ☐ | ☐ | ☐ | ☐ | Ⓗ | Experts disagree when their evaluation of the evidence leads them to defend different conclusions. Some conclusions are more reasonable and reflect a more comprehensive synthesis of the available information. |
| ☐ | ☐ | ☐ | ☐ | Ⓘ | Experts disagree because after examining and weighing the evidence they construct different interpretations about the safety of artificial sweeteners. They then evaluate these interpretations for their adequacy. |
| ☐ | ☐ | ☐ | ☐ | Ⓙ | Experts disagree about this issue because, like everyone else, they are confused about the safety of artificial sweeteners. So what they conclude is just their opinion. |

19. **Which three statements (A, B, C, etc.) from the list are most similar to your approach? Please indicate these in rank order below.**

Statement ◯ is most like how I think.

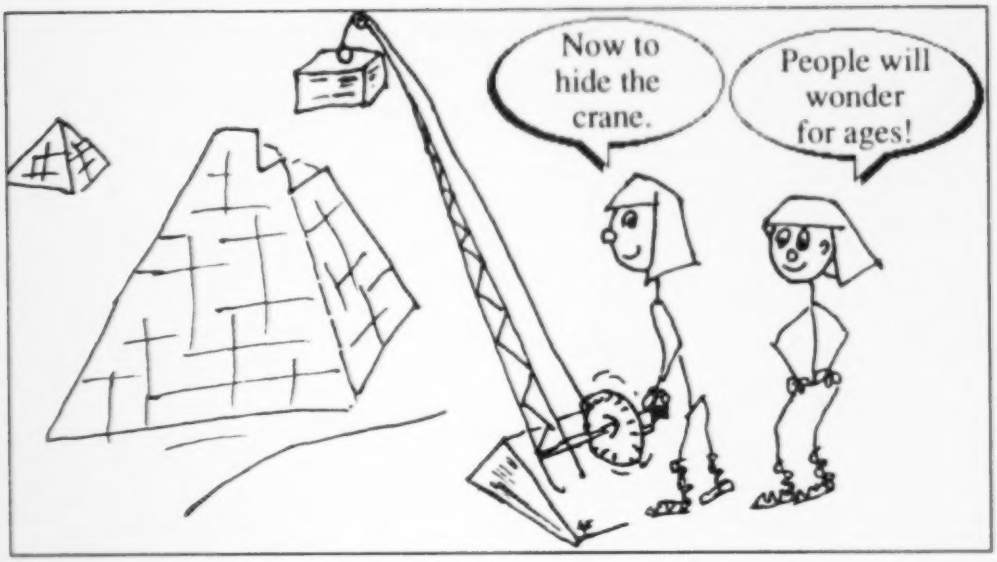
Statement ◯ is second most like how I think.

Statement ◯ is third most like how I think.

20. **How long did it take you to complete Issue A?** _____ minutes

**This completes Issue A.**
**Please DO NOT refer to your responses to Issue A as you complete Issue B.**

# Issue B




**Please note the current time:** _____

People have often wondered about things that have happened in the past. For example, many explanations have been offered about how the Egyptian pyramids were built. Some people claim that the pyramids were built as tombs for kings by the ancient Egyptians, using human labor, and aided by ropes, pulleys and rollers. Others have claimed that the Egyptians could not have built such huge structures by themselves, for they had neither the mathematical knowledge, the necessary tools, nor an adequate source of power.

**21. Which of the following responses is closest to <u>your</u> point of view about this issue?**

❑ I think the Egyptians built the pyramids by themselves.

❑ I do ***not*** think the Egyptians built the pyramids by themselves.

**22. People give different explanations for their point of view about how the pyramids were built. What is the basis for <u>your</u> point of view about this question?**

_____

_____

_____

_____

**23. How similar is each of the following statements to <u>the approach you used</u> to answer Question 22?**

| VS S D VD | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|
| ☐☐☐☐ Ⓐ | I heard in school that this is how the pyramids were built in Egypt, and I believe what I was told there. |
| ☐☐☐☐ Ⓑ | The facts aren't very clear because it happened so long ago and much of the evidence has been lost. So I just believe what seems right to me about how the pyramids were built. |
| ☐☐☐☐ Ⓒ | After comparing the interpretations on both sides of the issue, my point of view seems more reasonable to me because the evidence is stronger and the assumptions on which this view is based seem more valid. |
| ☐☐☐☐ Ⓓ | My point of view is based on my analysis of where the weight of the evidence lies. It is more probable because it best accounts for the evidence and other things I know about related topics, such as other ancient civilizations. |
| ☐☐☐☐ Ⓔ | There isn't much proof on either side of the issue about whether the Egyptians built the pyramids by themselves, so I believe what I want to believe. My point of view just makes sense. |
| ☐☐☐☐ Ⓕ | I look at the quality and density of the proof-claim of this issue and base my assumptions accordingly. The facts of this issue must be probabilistically migrated from that which is unproven to proven. |
| ☐☐☐☐ Ⓖ | Although there are several valid ways of looking at this issue, what I conclude is related to my assumptions about how historians do research and draw conclusions. These assumptions color how I interpret evidence. |

**24. Which three statements (A, B, C, etc.) from the list above are <u>most similar</u> to your approach? Please indicate these in rank order below.**

Statement ◯ is <u>most</u> like how I think.

Statement ◯ is <u>second</u> most like how I think.

Statement ◯ is <u>third</u> most like how I think.

25. **Some people think they already know for sure whether the Egyptians built the pyramids by themselves. Other people don't know for sure about this. What do you think? Why?** *(Write your answer in the space below.)*

_____

_____

_____

_____

26. ❑ I *don't* know for sure about this. ***[If you mark this answer, continue to Question 29, p. 15.]***

❑ I know for sure about this (or am fairly sure). ***[If you mark this answer, continue to Question 27, below.]***

27. **How similar is each of the following statements to the approach you used to answer Question 25?**

| VS | S | D | VD | | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | Ⓐ | I am sure about my opinion on whether the Egyptians built the pyramids by themselves because it just feels right to me and I just believe what I want to. |
| ☐ | ☐ | ☐ | ☐ | Ⓑ | I am sure about whether the Egyptians built the pyramids by themselves because books, television, or other sources of information describe how the Egyptians built the pyramids. |
| ☐ | ☐ | ☐ | ☐ | Ⓒ | I am sure for myself based on the facts I know, but other people may be just as sure of a different opinion based on the facts they know, and they have a right to their own opinion about this question. |
| ☐ | ☐ | ☐ | ☐ | Ⓓ | I am sure because my confidence impacts my single, infinite assessment of the underlying assumptions of this issue prior to collaborating data. |
| ☐ | ☐ | ☐ | ☐ | Ⓔ | I am sure because I have evaluated the evidence and its fit with related arguments and assumptions. As a result of that evaluation, I am confident about the validity of my conclusion. |
| ☐ | ☐ | ☐ | ☐ | Ⓕ | Knowing for sure is a pretty relative thing. It depends on your criteria for accepting an interpretation as certain. |
| ☐ | ☐ | ☐ | ☐ | Ⓖ | I am fairly sure because I have constructed a reasonable interpretation of the evidence, but I may not have interpreted the evidence adequately. |

28. **Which three statements (A, B, C, etc.) from the list above are most similar to your approach? Please indicate these in rank order below.**

Statement ◯ is most like how I think.

Statement ◯ is second most like how I think.

Statement ◯ is third most like how I think.

**If you completed Questions 27 & 28,**
**GO TO QUESTION 31 ON PAGE 16**

**29. How similar is each of the following statements to the approach you used to answer Question 25 (p. 14)?**

| VS | S | D | VD | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | Ⓐ I don't know for sure right now about whether the Egyptians built the pyramids by themselves because experts like archaeologists or historians don't yet have all of the information. But when they do, they will know for sure how they were built. |
| ☐ | ☐ | ☐ | ☐ | Ⓑ I don't know for sure because they were built too far in the past and the information has been lost. I do, however, have a personal opinion about this based on what I believe about the evidence. |
| ☐ | ☐ | ☐ | ☐ | Ⓒ I don't know for sure whether the Egyptians built the pyramids by themselves, but I could easily find out by simply asking someone who does know. |
| ☐ | ☐ | ☐ | ☐ | Ⓓ I don't know for sure because there are several valid interpretations of the evidence. However, after comparing these interpretations, I could construct a reasonable point of view based on its consistency with the available evidence. |
| ☐ | ☐ | ☐ | ☐ | Ⓔ I don't know for sure because I am not confident and because of the ocular opinions I hold about this point of view and the assumptions I can draw from its collusiveness. |
| ☐ | ☐ | ☐ | ☐ | Ⓕ I don't know for sure, but I think we can come very close to being sure. Even though there are competing interpretations of the evidence, I can conclude that one view is more probable because it best accounts for the evidence or because its underlying assumptions are more plausible. |
| ☐ | ☐ | ☐ | ☐ | Ⓖ I don't know for sure because people's interpretations of the evidence are always affected by their background, training, and assumptions. Thus, we have different ways of evaluating the evidence. |

**30. Which three statements (A, B, C, etc.) from the list above are most similar to your approach? Please indicate these in rank order below.**

Statement ◯ is most like how I think.

Statement ◯ is second most like how I think.

Statement ◯ is third most like how I think.

BEST COPY AVAILABLE

31. **When two people disagree about whether the Egyptians built the pyramids by themselves, is one person's point of view correct and other person's view incorrect? Why? Why not?** *(Write your answer in the space below.)*

32. ❑ No, one view is *not* correct (or mostly correct). ***[If you mark this answer, continue to Question 35, p. 17.]***

❑ Yes, one view is correct (or mostly correct). ***[If you mark this answer, continue to Question 33, below.]***

33. **How similar is each of the following statements to the approach you used to answer Question 31?**

VS S D VD **VS**=Very Similar; **S**=Similar; **D**=Dissimilar; **VD**=Very Dissimilar

| | | |
|---|---|---|
| ☐☐☐☐ | Ⓐ | Yes. One person would really know which point of view is correct about whether the Egyptians built the pyramids by themselves; the other person would just be wrong. |
| ☐☐☐☐ | Ⓑ | Yes. By correct, I mean that some explanations are very far along the continuum of what probably happened. It is a matter of choosing the position that seems most correct, most in line with the evidence, the best explanation. |
| ☐☐☐☐ | Ⓒ | Yes, one person's view is correct but we just don't know which one of these is correct right now. Someday I'm sure we will know which one is correct. |
| ☐☐☐☐ | Ⓓ | Yes, but not in an absolute sense. One point of view may be more sensible or more historically accurate than the other, but we can never know whether it is absolutely correct or not. |
| ☐☐☐☐ | Ⓔ | Yes, because the rule for illusiveness offers a solidified basis for choosing whether the Egyptians built the pyramids by themselves. |
| ☐☐☐☐ | Ⓕ | Yes, there is a correct answer, but it may not be knowable. We can only compare the interpretations and decide which seems more reasonable or accurate. |
| ☐☐☐☐ | Ⓖ | Yes, you could say that one point of view is correct, but that evaluation would be relative to a particular way of understanding whether the Egyptians built the pyramids themselves. |
| ☐☐☐☐ | Ⓗ | Yes, that person's point of view would be correct for that person. But for someone else, a different point of view would be correct. |

34. **Which three statements (A, B, C, etc.) from the list above are most similar to your approach? Please indicate these in rank order below.**

Statement ◯ is most like how I think.

Statement ◯ is second most like how I think.

Statement ◯ is third most like how I think.

BEST COPY AVAILABLE

**If you completed Questions 33 & 34, GO TO QUESTION 37 ON PAGE 18**

**35. How similar is each of the following statements to <u>the approach you used</u> to answer Question 31 (p. 16)?**

| VS S D VD | | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|---|
| ☐☐☐☐ | Ⓐ | No, you can't say one is correct and one is incorrect because there is no definite proof; they're both just opinions about the facts. What each person thinks is an individual thing. |
| ☐☐☐☐ | Ⓑ | No, not until one is proven. Until it is proven, it is just a matter of what you want to believe about whether the Egyptians built the pyramids by themselves. |
| ☐☐☐☐ | Ⓒ | No, not in an absolute sense. One point of view may be more feasible or more historically accurate than the other, but we can never know absolutely whether or not it is correct. |
| ☐☐☐☐ | Ⓓ | No, you cannot say one point of view is correct because a person's evaluation is relative to a particular way of understanding whether the Egyptians built the pyramids by themselves. |
| ☐☐☐☐ | Ⓔ | No, I cannot say absolutely that one view is correct because the evidence is incomplete and always will be, but I can evaluate one point of view as being a more reasonable interpretation than another. |
| ☐☐☐☐ | Ⓕ | No, because we can never have all the evidence or know if the information is being interpreted correctly. Nevertheless, we can judge one point of view as more plausible given what we now know. |
| ☐☐☐☐ | Ⓖ | No, because being correct requires interpreting how well misinformed your facts and other criteria are for being knowledgeable about whether the Egyptians built them by themselves or not. |
| ☐☐☐☐ | Ⓗ | No, because we can't say which point of view is correct. It is so hard to prove how the pyramids were built when it happened so long ago. So much evidence has been lost and we can't go back in time to find out how it really happened. |
| ☐☐☐☐ | Ⓘ | No, because there isn't that much proof about whether the Egyptians built them by themselves or not, so either view could be right or wrong. Until they prove it, you can't say which is correct. |

**36. Which three statements (A, B, C, etc.) from the list above are <u>most similar</u> to your approach? Please indicate these in rank order below.**

Statement ◯ is <u>most</u> like how I think.

Statement ◯ is <u>second</u> most like how I think.

Statement ◯ is <u>third</u> most like how I think.

**37. How is it possible that two experts in the field who have the same information can genuinely disagree on this issue and arrive at different conclusions?** *(Write your answer in the space below.)*

**38. How similar is each of the following statements to the approach you used to answer Question 37?**

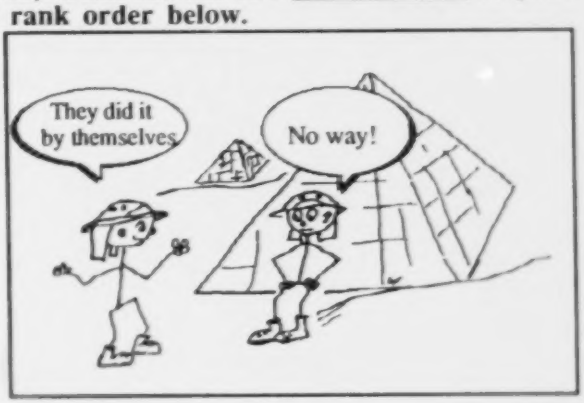
| VS S D VD | | VS=Very Similar; S=Similar; D=Dissimilar; VD=Very Dissimilar |
|---|---|---|
| ☐☐☐☐ | Ⓐ | Experts disagree because they approach the problem with different conclusions already in mind and find evidence to support their conclusions. |
| ☐☐☐☐ | Ⓑ | Experts disagree because their beliefs are relative to their own perspective. As a result they interpret the evidence differently. |
| ☐☐☐☐ | Ⓒ | Real experts who are honest will not disagree. |
| ☐☐☐☐ | Ⓓ | Experts disagree because the contractual hard evidence is synthesized into available belief systems about different comprehensive factual analyses. |
| ☐☐☐☐ | Ⓔ | Experts disagree because it is not yet known how it was done. Until there is more evidence, they will believe what they want to believe about it. |
| ☐☐☐☐ | Ⓕ | Experts disagree because they begin with different beliefs and experiences. These lead them to look at the facts differently. |
| ☐☐☐☐ | Ⓖ | Experts disagree when their synthesis of the available information leads them to defend different conclusions. Some of these conclusions are more reasonable or plausible and reflect a more comprehensive synthesis of the available information. |
| ☐☐☐☐ | Ⓗ | Experts disagree about whether the Egyptians built the pyramids by themselves because like everyone else, they are confused about how the pyramids were built. So what they conclude is just their opinion. |
| ☐☐☐☐ | Ⓘ | Experts disagree because after examining the evidence and interpretations, they differ about what is the more reasonable conclusion. |
| ☐☐☐☐ | Ⓙ | Experts disagree as a result of differences in assumptions, emphasis, and methods of evaluating and interpreting evidence. However, the adequacy of their conclusions can also be evaluated. |

**39. Which three statements (A, B, C, etc.) from the list are most similar to your approach? Please indicate these in rank order below.**

Statement ◯ is most like how I think.

Statement ◯ is second most like how I think.

Statement ◯ is third most like how I think.




**40. How long did it take you to complete Issue B?** _____ minutes

BEST COPY AVAILABLE

**THANK YOU VERY MUCH!**
We are interested in hearing any reactions or comments you have about completing this questionnaire. *Please use the space below to offer your comments.*

In what ways have your educational experiences contributed to your thinking about issues like these?

APPENDIX B

# Technical Manual to Accompany the Reflective Thinking Appraisal (RTA; Ver. 1.0)

Phillip K. Wood[1]

Patricia King

Karen S. Kitchener

Cindy Lynch

©1994 Reflective Thinking Associates

All Rights Reserved.

1. Address Queries concerning this manual to Phillip Wood, 210 McAlester Hall, University of Missouri-Columbia, Columbia, MO 65211


©1994 Reflective Thinking Associates

Contents

Preface ..... 3
Administration of the RTA ..... 5
RTA Instrument Design and Scoring ..... 6
Subjects ..... 7
Scoring ..... 11
Sequentiality ..... 16
Educational Level Differences ..... 19
Sex Differences ..... 22
Summary ..... 23
References ..... 25

Tables

Table 1 Sample Sizes by Site and Educational Level ..... 8
Table 2 Age Means by Site and Educational Level ..... 10
Table 3 Coefficient Alpha Internal Consistency Estimates
by Educational Level & Site ..... 15
Table 4. Observed and Expected Response Pattern Frequencies for
Predominant and Minor Stages ..... 18


©1994 Reflective Thinking Associates

## Preface

This manual describes administration procedures and technical information concerning the Reflective Thinking Appraisal (RTA), a measure of ill-structured problem solving based on the Reflective Judgment Interview (RJI, Kitchener & King, 1981).

### Instrument History:

The RTA has gone through four revisions during the course of its development. The version of the RTA presented in this manual represents the version for which the most extensive data were available. This history is presented in order to inform the reader of changes which were made during the course of the development of the instrument and the data on which these decisions were based. Note that the most current version of the RTA is not the same as the norm sample outlined here. Minor changes in wording and a few response possibilities were added as a result of subject comments for two questions in the most recent version. It seems reasonable to assume that the general psychometric properties and levels of performance will be quite similar to those reported here.

Version 0.1 (1991) First version of the RTA developed based on the Pyramids dilemma

Version 0.2 (May-Sept. 1992) Food Additives Dilemma added instrument piloted on 165 students- initial psychometric properties investigated pilot data consisted largely of early undergraduate subjects. Additional 132 subjects assessed in September. Final data set for version 0.2 consisted of 56 freshmen, 92 Sophomores, 142 seniors, and 44 graduate students (composed mostly of entering doctoral students),. Item analyses and psychometric information used to identify items which were ambiguous for students and refine item behavior.


©1994 Reflective Thinking Associates

Version 0.3 (Sept. 1992-June 1994) Minor refinements to items made. Data administered to a pilot norming sample of 534 individuals (described in more detail in the Subjects Section below).

Version 1.0 (June 1994-present). Minor changes made to wording of some items. Two items with low endorsement rates dropped, changes made to some item stems based on item analyses of the pilot data. The instrument is available in two forms (Forms P and C) in order to enable researchers to counterbalance the order of presentation of the Pyramids and Saccharin dilemmas.


©1994 Reflective Thinking Associates

## Administration of the RTA.

### Appropriate Subject populations

Like the Reflective Judgment Interview, the RTA is designed to assess reasoning about ill-structured problems for which no single correct answer exists. As such, use of the instrument with young populations (e.g., junior high school students) is not advisable, given both the complex nature of the task and the complexities involved in completing the instrument. Given the highly verbal nature of the instrument, use of the RTA may be a problem for some minority groups or individuals whose first language is not English.

### Instructions & Timing.

Although the RTA can be administered to a large group of subjects at the same time, it is essential that these test administrations be proctored in order to assure that subjects questions concerning the instrument are answered and to assure that no gross violations of testing procedure occur (such as subjects copying others' responses, discussing the issues during the assessment, or pressuring an individual to complete the instrument quickly). Researchers should generally allow about an hour to complete the measure, although many subjects finish in 30 to 40 minutes.

### Instructions to Subjects

The proctor should make the following points to individuals prior to administering the test instrument:

(1.) This instrument is designed to assess how people reason about issues which do not have a single correct answer. Different individuals have different opinions about these matters.

(2.) We are asking individuals to write out their opinions about a number of questions and then to ask them to match statements taken from interviews to their own opinion. In many cases, these statements may not be exact matches to their own opinions, and subjects will have to match the best available statement.


©1994 Reflective Thinking Associates

(3.) The time allowed for the RTA is usually ample for everyone to complete it. Subjects should consider each question carefully, but should also pace themselves so that they can finish in an hour.

(4.) Students should assign their educational level based on their current status, and not their highest level of completed education. (E.g., a college freshman in his/her first semester should indicate that they are a college freshman and not a high school senior).

(5.) Some of the statements in the instrument may not make logical sense or may appear to be typographical errors. They are not. These items are designed to identify individuals who wish to sound sophisticated or educated but who do not consider the meaning of an item. Subjects are instructed to mark these items as "Very Dissimilar" to their own opinions.

(6.) If, during the assessment, subjects have any questions, the proctor is available to answer questions. During assessments of undergraduate students it has been our experience that these questions generally take the form of not understanding the definition of a particular word (e.g., "premeditated" or "synthesized"). Questions from graduate populations usually concern the fact that a particular "meaningless" item does, in fact, make no sense. If the individual does not understand the meaning of a word, the proctor should give her/him a dictionary definition. If the query concerns the problem of an item making no sense, remind the individual that items which do not make sense should be rated as very dissimilar to their own opinion.

RTA Instrument Design and Scoring.

The RTA in its present form consists of two topics. The first topic concerns an issue of historical interpretation; specifically whether the ancient Egyptians were capable of building the pyramids by themselves or required assistance The second topic deals with scientific inference under uncertainty specifically as it relates to the


©1994 Reflective Thinking Associates

issue of the safety of artificial sweeteners. For each topic four questions are designed to assess: (1.) The justification for their opinion; (2.) The degree to which subjects feel their opinion is true with certainty; (3.) Subject views as to whether and how different opinions on the topic may be judged correct or incorrect; (4.) Subject views as to how experts could disagree about the topic.

For Questions 2 and 3, two possible sets of stems are available to the subject. In Question 2 depending on whether they believe that their opinions are known with certainty. In Question 3, two possible sets of items are possible depending on whether subjects believe that opinions can be judged as correct/incorrect. For each of the possible statements in the instrument, subjects are asked to rate, on a four-point scale, the degree to which a given statement is similar to their own opinion. Following that, subjects are asked to consider their responses and to rank the three statements in terms of their similarity to the subject's own opinion.

## Subjects

Data from 534 subjects are included in the norming sample for this study. Predominantly these data were taken from assessment sites at Bowling Green State University (N=91), Denver University (N=42), and the University of Missouri-Columbia (N=395). One set of subjects from the University of Missouri site were taken from students enrolled in Introductory Psychology. These individuals reported that they were freshmen, sophomores, or juniors. Seniors from this site were taken from an end-of-year senior assessment of psychology majors. Finally, three individuals were taken from Institution coded as 9 and another three were from an Institution coded as 99. Sample Sizes overall and by each site are given in Table 1.


©1994 Reflective Thinking Associates

Table 1.

Sample Sizes for Norming Sample by Site and Educational Level

| Educational Level | N | Inst. 1 (BGSU) | Inst. 2 (DU) | Inst. 3 (MO) | Other -- |
|---|---|---|---|---|---|
| Undergraduate | | | | | |
| Freshmen | 188 | 23 | 3 | 162 | |
| Sophomore | 65 | 26 | 2 | 37 | |
| Junior | 40 | 18 | 6 | 16 | |
| Senior | 114 | 14 | 7 | 93 | |
| Graduate | | | | | |
| Currently completing: | | | | | |
| 1st half of master's course work | 20 | | 4 | 16 | |
| 2nd half of master's course work | 21 | 0 | 1 | 20 | 2 |
| 1st half doctoral course work | 21 | 1 | 3 | 15 | 2 |
| 2nd half doctoral course work | 36 | 5 | 10 | 19 | 2 |
| Doctoral Coursework Completed | 29 | 4 | 6 | 17 | |
| Total | 534 | 91 | 42 | 395 | |





RTA Technical Manual June 6, 1994 2:02 pm Page 8
©1994 Reflective Thinking Associates

As can be seen from this table, the design of the norm sample is somewhat confounded across site. Data come primarily from the BGSU and MO samples. More females than males were present in the study (63.11% female), with relatively equal numbers of men and women occurring only in the 2nd Half of Doctoral Coursework and Doctoral Coursework Completed groups. This imbalance in the study is discussed under the analysis of sex effects discussed below. On the average, these individuals represented traditionally aged college students, with the possible exception of the undergraduate individuals from the University of Denver (Table 2) where the freshmen and sophomore individuals were two years older than their counterparts at the other two sites.

It should also be noted that in previous pilot work with the instrument that undergraduates sometimes misreport their educational level. Specifically, students tend to report their status based on the number of years in attendence at the university and not based on the number of credit hours they have attended. Stated age for these subjects was calculated by computing the difference between the last two digits of the year of their birth from the year that the instrument was taken. This calculation is also slightly inaccurate, in that the ages reported may vary up to a year from the actual chronological age of the individual.


©1994 Reflective Thinking Associates

Table 2.

Age Means for Norming Sample by Site and Educational Level

| Educational Level | Overall | Inst. 1 (BGSU) | Inst. 2 (DU) | Inst. 3 (MO) | Other -- |
|---|---|---|---|---|---|
| Undergraduate | | | | | |
| Freshmen | 18.5 | 18.3 | 20.7 | 18.5 | |
| Sophomore | 19.8 | 19.9 | 21.5 | 19.7 | |
| Junior | 20.2 | 20.2 | 20.3 | 20.2 | |
| Senior | 22.8 | 23.6 | 25.9 | 22.4 | |
| Graduate | | | | | |
| Currently completing: | | | | | |
| 1st half of master's course work | 26.7 | - | 27.8 | 26.4 | |
| 2nd half of master's course work | 28.5 | - | 26.0 | 28.7 | 38.5 |
| 1st half doctoral course work | 31.3 | 39.0 | 38.7 | 28.3 | 31.0 |
| 2nd half doctoral course work | 32.0 | 27.5 | 36 | 31.9 | 38.0 |
| Doctoral Coursework Completed | 33.1 | 31.8 | 37.3 | 31.3 | |
| Total | 22.7 | 29.6 | 22.1 | 21.2 | |

RTA Technical Manual June 6, 1994 2:02 pm Page 10
©1994 Reflective Thinking Associates

## Scoring.

### Stage Utilization Scores.

Current scoring of the RTA closely follows the general scoring scheme of Rest's (1979) Defining Issues Test (DIT), with some minor modifications. Like the DIT, composite scores for the instrument are only based on the ranking information of the instrument, and not on the similarity scores associated with each stem. Like the DIT, the scoring scheme for the RTA first produces a raw percent stage utilization score for each question based on the rankings of statements for the question. The stage corresponding to the most highly ranked statement is assigned a weight of .5, the stage corresponding to the second ranked statement is assigned a weight of .3, and the stage associated with the third ranking is assigned a weight of .2. These weights are then averaged across statements to arrive at stage utilization scores for stages 2 through 7 as well as the meaningless items. For example, if an individual ranked statements corresponding to levels 4, 3, and Meaningless, that individual would receive a stage utilization score of .5 for Stage 4, .3 for Stage 3, and .2 for Meaningless.

### Percent Stage Utilization Scores.

Since some individuals do not assign all three rankings for a given question and since the utilization scores do not convey the relative percent of stage utilization (since the Meaningless items do not correspond to a developmental level in the scheme). Utilization scores are then converted to percent stage utilization scores by dividing stage utilization scores by the scores which were awarded a stage score for a question. For example, since the sample ratings above included a meaningless response, the percent stage utilization scores for this individual are calculated as: .5/.8=.625 for Stage 4, and .3/.8=.375 for Stage 3.

As a result, percent stage utilization scores for a given topic (such as the Pyramids topic) are then computed as an average of the percent stage utilization scores


©1994 Reflective Thinking Associates

across the four questions for that topic. As opposed to the RJI, where percent stage utilization scores have been shown to correlate substantially across dilemmas, the percent stage utilization scores of the RTA do not appear to correlate well across the two topics. Correlations of the percent stage utilization scores across the two topics based on all available data ranged from .25-.36 for stages 2-6 and were only .57 for Stage 7.

Overall Composite Score.

A variety of scoring schemes have been explored based on both the initial pilot and this normative sample, including reverse weighting schemes based on stage scores, percent of responses across the higher levels of the model, as well as various combinations of particular stage scores. The overall score for the measure proposed here is that which has proved most internally consistent across the two topics and also that composite which produced the most pronounced educational level effects. If the percent stage utilization scores for all stages are assumed to be marginal proportions from a multinomial distribution and if the stage levels of the Reflective Judgment model represent roughly interval levels, the marginal mean of this Dirichlet process can be computed as: $\sum_{i=2-7} i \cdot p_i$ where i indexes the possible stage scores and $p_i$ indicates the percent stage utilization score for a given level. As such, this overall score represents a continuous level measure. Given the fact that the resulting scores fall on the interval 2-7, it is appropriate to note that the actual correspondence of these scores with those obtained by the RJI is the object of ongoing research. Scores can only be interpreted relative to the educational level differences found in a given study or the normative samples reported here.

The internal consistency of the overall score was calculated for the normative


©1994 Reflective Thinking Associates

sample by educational level and for those subsamples where sufficient sample size was available. Some caution regarding the magnitude of the obtained internal consistency estimates for the educational level samples by site, due to the small sample sizes (See Table 1). Internal consistency estimates were calculated separately based on all available response as well as those individuals who did not exceed the criterion for excessive endorsement of meaningless items (discussed below). As can be seen, the RTA appears to be more internally consistent for the graduate populations than for the undergraduate populations. It should be noted that the Bowling Green undergraduate samples were much smaller than their University of Missouri-Columbia counterparts. This probably accounts for the apparent larger coefficient alphas for this site, since coefficient alpha is dependent on sample size.

Consistency Check-Meaningless Response

Like the DIT, the RTA also contains a number of items which are designed to assess subjects' tendency to endorse statements which contain impressive vocabulary or on the basis of their pretentiousness rather than their meaning. In the RTA, the meaningless score is calculated the same as the stage utilization scores outlined above. Analyses based on both the pilot work conducted on version 0.2 and the normative sample described here suggest that a cutoff score of .05 has resulted in slightly higher measures of internal consistency (See Table 3) and in making the patterns of means across educational level (discussed below) more distinct. For the normative sample as a whole, the sm cutoff score represents roughly the 75th percentile of the distribution of scores as a whole. Individuals with high meaningless scores tended to be a function of educational level. For the Denver University site, five of the eleven undergraduates who were not yet seniors scoring higher than the cutoff. for the University of Missouri-Columbia site, seven of the 16 undergraduate juniors scored higher than this cutoff. For all remaining groups, the percent of individuals scoring higher than the cutoff was equal to or less than 25%.


©1994 Reflective Thinking Associates

Taken together, this pattern of inconsistency suggests two limitations of the RTA which bear further study. It is possible that some individuals may wish to "sound educated" at the expense of providing scoreable information (such as the individuals at the Denver site who scored above the cutoff). Additionally, as has been found for the DIT (Rest, 1979) the motivation of students may also affect their endorsement rates for meaningless items. The juniors who were assessed at the University of Missouri-Columbia seem to fall into this category in that the internal consistency for this group jumps markedly if individuals with high rates of meaningless response are excluded and also in light of the consideration that these individuals consisted of college juniors who were enrolled in Introductory Psychology. Generally, these students consisted of individuals majoring in other disciplines who are merely taking the Introductory Psychology course in order to secure enough credits to register in a given semester.


©1994 Reflective Thinking Associates

Table 3

Coefficient Alpha Internal Consistency Estimates by Educational Level & Site[2]

| Educational Level | Overall | Inst. 1 (BGSU) | Inst. 2 (DU) | Inst. 3 (MO) | Other -- |
|---|---|---|---|---|---|
| Undergraduate | | | | | |
| Freshmen | .62 (.68) | .88 (.87) | | .48 (.59) | |
| Sophomore | .68 (.76) | .58 (.62) | | .70 (.80) | |
| Junior | .63 (.61) | .83 (.48) | | .35 (.83) | |
| Senior | .71 (.70) | .83 (.82) | | .67 (.63) | |
| Total Undergraduate | .67 (.70) | .76 (.75) | | .60 (.67) | |
| Graduate | | | | | |
| Currently completing: | | | | | |
| 1st half of master's course work | .83 (.85) | | | .79 (.83) | |
| 2nd half of master's course work | .73 (.75) | | | .78 (.75) | |
| 1st half doctoral course work | .76 (.71) | | | .84 (.77) | |
| 2nd half doctoral course work | .64 (.55) | | | .57 (.47) | |
| Doctoral Coursework Completed | .82 (.83) | | | .86 (.90) | |
| Total Graduate[3] | .76 (.74) | | | .78 (.75) | |
| Grand Total | .77 (.79) | | | | |

2. Numbers in parentheses indicate estimates derived when individuals with high rates of meaningless statement endorsement are excluded

3. Excluding Individuals with Doctoral Coursework Completed


©1994 Reflective Thinking Associates

## Sequentiality

Since the RTA contains separate estimates of the percent stage utilization scores for levels 2 through 7 of the scheme, it is also possible to assess the "cross-sectional" sequentiality of the instrument: the degree to which those level most highly ranked by subjects are adjacent to the stage immediately preceding or following it in the postulated developmental sequence. Davison (1977) proposed a statistical test for the presence of such developmental sequences in developmental data. Davison et al.'s (19XX) examination of several cognitive developmental instruments found that only the DIT demonstrated such sequentiality. Davison's (1977) analysis of RJI data as well as Wood's (1994) analysis of extant RJI data sets found that the RJI demonstrated this type of sequentiality as well, although Wood's analysis of available RJI data found that the degree of this sequentiality varied as a function of developmental level, with higher stages of Reflective Judgment showing much more variability across stages than earlier stages.

For the data of this pilot study, the predominant level of response was defined as that stage for which the percent stage utilization score was highest. In the cases of ties, the predominant level was assigned randomly among available responses (this occurred in less than 99% of the data). The subdominant level was defined as the next most frequently used level and ties were, again, broken randomly.

Table 4 shows the observed predominant and subdominant stages for the data from the normative sample. The data shown here represent individuals who did not exceed the cutoff for meaningless item endorsement described above. The patterns of statistical significance described here, however, were no different for the data set as a whole. Davison's (1977) test proceeds by constructing a cross-tabulation table of predominant by subdominant responses. The diagonal elements of this cross-tabulation table are defined as structural zeroes, and the patterns of response across individuals is examined to see if it significantly departs from a null model of independence. For these data, the $\chi^2$ test statistic was highly significant ($\chi^2(df=19)=96.21$; $p<.0001$), thus the observed cross-tabulation patterns observed in Table 4 do not appear to be due to random variability. Next, Davison proposes a test of sequentiality which allows an additional probability mass to be added to sequential cells in the table. This additional probability is assumed to be the same for subdominant scores across all stages. The fit of Davison's sequential model to these data, while significantly reducing the magnitude of the $\chi^2$ test statistic, still yields an unacceptably high value ($\chi^2$ $(df=18)=29.80$; $p=.039$). As such, this result paralleled Wood's (1994) analysis of sequentiality in the RJI, where it was found that Davison's model failed to recapture the observed sequentiality patterns in the data since it assumed that the magnitude of predominant/subdominant response was a same for all levels of the scheme. According-


©1994 Reflective Thinking Associates

ly two additional models were tested: In the first alternative model, it was assumed that the probability mass associated with subdominant response was the same within each level of the scheme, but that these probability masses varied across the dominant levels. This model did not fit the data appreciably better than Davison's original sequentiality model ($\chi^2$(df=13)=28.57; p=.007). Finally, final extension of the sequentiality model allowed the magnitude of sequential response to vary as a function of dominant level and further allowed the probability mass associated with sequential response to vary as a function of whether a subdominant stage was higher or lower than the dominant stage. This model fit the data quite well ($\chi^2$ (df=9)=3.20; p=.49).

As such, explorations of the sequentiality of the RTA reveal that documents an internally sequential progression, much like that found in the RJI. However, like the RJI, the patterns of sequential response appear to differ as a function of both the dominant stage and whether that stage is higher or lower than the present stage. For example, for some dominant stages adjacent scores higher than the dominant level are observed more frequently (i.e., for individuals with a dominant score of 6 in Table 4, a Stage 7 subdominant score is more likely than a subdominant score of 5: 33 individuals scored at level 7, while only 4 individuals scored at Level 4 for this group). On the other hand, for other stages lower stages are more likely than higher ones. For example individuals with a predominant score of Stage 5 are more likely to show a subdominant score of Stage 4 than they are of Stage 6. Other stages (such as Stage 4) appear to be equally divided between higher and lower subdominant responses. Dominant stage responses of Stage 2 are quite infrequent, probably due to the administration of the instrument to college populations.


©1994 Reflective Thinking Associates

Table 4
Observed and Expected Response Pattern Frequencies for Predominant and Minor Stages

| Predominant Stage | Minor Stage 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 2 | -[4] | 0[5] | 1 | 0 | 1 | 1 |
| | | .38[6] | .74 | .37 | 1.01 | .51 |
| | | .58[7] | .90 | .19 | .88 | .45 |
| | | .00[8] | 1.04 | .20 | 1.17 | .59 |
| | | .00[9] | 1.03 | .52 | 1.01 | .44 |
| 3 | 0 | - | 12 | 3 | 2 | 4 |
| | .32 | | 5.86 | 2.89 | 7.93 | 4.00 |
| | .65 | | 12.67 | .94 | 4.44 | 2.29 |
| | .64 | | 11.36 | .92 | 5.37 | 2.72 |
| | .00 | | 12. | 2.37 | 4.64 | 1.99 |
| 4 | 4 | 55 | - | 47 | 84 | 33 |
| | 3.90 | 36.59 | | 35.55 | 97.70 | 49.26 |
| | 4.19 | 52.59 | | 46.40 | 79.08 | 40.74 |
| | 3.90 | 53.68 | | 48.32 | 77.75 | 39.35 |
| | 4.71 | 55 | | 47 | 81.35 | 34.95 |
| 5 | 0 | 1 | 6 | - | 1 | 0 |
| | .12 | 1.13 | 2.22 | | 3.01 | 1.52 |
| | .07 | .30 | 3.55 | | 3.44 | .64 |
| | .07 | .25 | 3.29 | | 3.71 | .68 |
| | .08 | .29 | 6 | | 1 | .62 |
| 6 | 2 | 4 | 20 | 4 | - | 33 |
| | 1.25 | 11.67 | 23.03 | 11.34 | | 15.71 |
| | 1.00 | 4.54 | 19.50 | 11.08 | | 26.87 |
| | 1.26 | 4.38 | 20.46 | 9.34 | | 27.66 |
| | 1.17 | 4.10 | 20.73 | 4 | | 33 |
| 7 | 1 | 3 | 19 | 9 | 57 | - |
| | 1.41 | 13.24 | 26.13 | 12.86 | 35.35 | |
| | 1.10 | 4.98 | 21.37 | 4.39 | 57.16 | |
| | 1.24 | 4.69 | 21.86 | 4.22 | 57.00 | |
| | 1.03 | 3.61 | 18.24 | 9.12 | 57 | |

4. Dashes indicate response patterns (cells) which cannot occur.
5. The First number in each cell represents the observed frequency.
6. Second numbers indicate predicted cell frequencies under quasi-independence model.
7. Third numbers indicate predicted cell frequencies under Davison's original sequentiality model.
8. Fourth numbers indicate predicted cell frequencies under sequentiality model allowing each dominant stage to have different degrees of sequentiality.
9. Fifth numbers indicate predicted cell frequencies under sequentiality model allowing each dominant stage to have different degrees of sequentiality at higher and lower adjacent stages.


©1994 Reflective Thinking Associates

Educational Level Differences in Performance.

When the data from the entire study are analyzed to test if differences in overall score vary as a function of educational level, the overall model reveals that significant differences exist across educational levels ($F(8.398)=22.03$; $p<.001$). Follow-up Waller-Duncan tests of the differences between the educational levels revealed that undergraduates scored lower on the RTA (Means ranging from 4.69-4.91) than graduates and that master's level students in their first half of the coursework (mean 5.18) did not score significantly different that master's level students in their second half (Mean=5.38), but did score lower than all other graduate educational levels (Mean 5.46-5.59). A dot plot of these scores by educational level is given in Figure 1. In this figure, a circle represents the mean for each educational level, with an error bar showing the magnitude of the 95% confidence region of the error measurement. Vertical bars indicate those educational levels which were not statistically different under the Waller-Duncan test.

The overall general linear model establishes that the RTA can successfully grossly differentiate undergraduate from graduate performance. Often, however, researchers interested in educational evaluations would not wish to compare the performance of graduate students relative to undergraduates. Previous research using the RJI has also found that the variability in performance in Reflective Judgment varies as a function of educational level, with lower educational levels demonstrating significantly less variability in performance than higher educational levels. (Wood, 1994b). Note also that the design of the study is extremely unbalanced, with larger numbers of students in the senior and freshman groups, and comparatively fewer graduate students than undergraduates. The effect of unbalanced sample sizes and differential variability in performance is to reduce the statistical power of hypothesis testing in the general linear model, causing the researcher to falsely fail to reject the null hypothesis. For these conceptual and


©1994 Reflective Thinking Associates

statistical reasons, additional general linear models were specified which examined whether educational level differences separately for the undergraduate and graduate populations. A general linear model based on only undergraduate data revealed that statistically significant differences exist between educational levels ($F(3,289)=5.04$; $p<.01$). Waller-Duncan contrasts conducted at the .05 level revealed that freshmen (Mean=4.67) scored significantly lower than seniors (Mean=4.92). No other differences were found between groups. These means and this pattern of differences across groups is indicated in Figure 2. An analogous general linear model for the graduate students (based on only individuals who had not yet completed their coursework) revealed a statistically significant overall model ($F(3,75)=3.00$; $p=.035$) and the same pattern of group differences as found in Figure 1. It should also be noted that those individuals who had indicated that they had completed their coursework toward the Ph.D. were more variable in performance than other graduate education levels. The overall model was not statistically significant if individuals who had completed their doctoral coursework were included ($F(4,97)=2.09$; $p=.09$).

Although it appears likely that statistically significant differences exist across levels of undergraduate study, it is worth noting that the relative magnitude of the effect size associated with these differences is much smaller than that associated with the RJI. Recall that effect size is a statistical index of the amount of difference, in standard score units across groups. Pascarella & Terenzini (1991) calculate a rough estimate of the effect size for a number of educational outcomes associated with year of study as (freshman year mean-senior mean)/standard deviation of freshman year and report that the effect size associated with the RJI is approximately 1. (Although note that Wood's (1994) re-analysis of available RJI data revealed substantial differences in mean score as a function of sample and institution.) The effect size found in this pilot study for the undergraduate data was


©1994 Reflective Thinking Associates

only .48. The effect size across graduate samples was slightly larger (.66). At this stage, it is difficult to determine the cause of this discrepancy in effect sizes. It is possible that this is due to the fact that many RJI studies involve carefully matched or longitudinal assessments, while the pilot data gathered here was based mainly on intact available groups. It is also possible that the discrepancy in effect sizes is due to the fact that the RTA is a recognition task, whereas the RJI is a production task. Finally, given the more highly reliable nature of the RJI, it is possible that the distribution of freshman scores on the RTA is more contaminated with measurement error.

Site Differences Within Educational Level

Finally, institutional evaluations often involve the assessment of differences in performance across educational settings. As such, it is of interest to know if the RTA can be used to differentiate performance across institutions and if different patterns of performance obtain in these institutions across educational levels. Examination of this question based on the available data is difficult, given the small number of sites and the unbalanced nature of the data. A preliminary examination of this question, though, seemed possible, based on the sizes of the Bowling Green and University of Missouri undergraduate samples. A general linear model examining the effects of educational level (freshman through senior) and site (Bowling Green and University of Missouri) revealed statistically significant overall model ($F(7,285)=4.08$; $p<.001$) and a statistically significant effect for year of study ($F(3,285)=5.19$; $p=.0017$) but not for site ($F(1,285)=2.66$; $p=.1042$). Surprisingly, a statistically significant interaction between educational level and site was found ($F(3,285)=3.43$; $p=.0175$). If order to further probe this interaction, two additional general linear models were run examining educational level patterns within each site. For the University of Missouri data, a significant main effect for educational level was found ($F(3,228)=8.14$; $p<.0001$), with seniors (Mean=4.96) scoring


©1994 Reflective Thinking Associates

higher than the other three educational levels (Means ranging from 4.61-4.70). For the Bowling Green data, however, no statistically significant differences across educational level were found ($F(3,57)<1$). The mean score for freshman (4.68) and the mean for the seniors (4.90) was similar to counterparts at the University of Missouri, however, the performance of the sophomores and juniors seemed much higher (4.84 and 4.92, respectively).[10] Given the interaction of site with educational level the replication of educational level differences across multiple educational institutions seems appropriate before strong claims regarding the ability of the RTA to identify educational level differences in reasoning about ill-structured problems.

Sex Differences.

Although a preliminary analysis of overall score as a function of sex favoring males, no analyses based on undergraduate data only, graduate data only, or all available data were able to identify a statistically significant effect for sex. Due to the unbalanced nature of the design (more women than men were tested) and the confounding of educational level with sex, no statement regarding the presence of sex-related differences in performance on the RTA seems appropriate at this time. The differences in educational level remained after controlling for gender, however ($F(3,296)=3.23$; $p<.001$). For the graduate data, gender and educational level were confounded: Three men and fourteen women indicated that they were in the first half of heir master's program; Five men and twelve women indicated that they were in the first half of their doctoral studies. By contrast, thirteen men and 10 women indicated that they had completed their coursework for the doctorate. Provisionally, a test of whether educational level differences exist for both female and male graduate students by conducting separate general linear models testing whether educational level effects existed. For the women, the analysis showed a significant

10. Note, though, that a t-test of the freshman means from the two institutions is statistically significant ($t(140)=2.22$)


©1994 Reflective Thinking Associates

main effect for educational level ($F(4,58)=3.24$; $p=.0182$). Waller-Duncan a posteriori contrasts at the .05 level indicated that female graduate students in their first half of master's degree coursework (Mean=5.0221) were not significantly different from those who indicated that they had completed their doctoral coursework (mean=5.25), but did differ from all remaining groups. (Mean=5.61, 5.55, and 5.49 for the students in the 2nd half of their master's degree coursework, first half of doctoral coursework, and second half of doctoral coursework, respectively.) For male students, no statistically significant differences were found ($F(4,34)=2.40$; $p=.0691$). Patterns of mean scores for the men by educational level were somewhat similar, however.

## Summary

Taken together, the results of the analysis of the RTA data to date suggest it appears to document differences in reasoning about ill-structured problems which are similar to those which have been found using the RJI. Although differences between undergraduate and graduate students were found using the instrument, similar patterns of educational level performance were not found for two sites, Bowling Green State University and the University of Missouri-Columbia. To some extent, this failure may be due to the fact that the subjects taken from the University of Missouri were, with few exceptions, drawn from the same department while those from Bowling Green State University represented a more heterogenous (and possibly confounded) population.

Like the DIT, the RTA is also susceptible to the demand characteristics of the situation and/or the motivational level of the student. To this end, it appears that the meaningless response consistency checks serve to control for the tendency of some subjects to endorse pretentious-sounding responses. This control is by no means complete, however, and future work is required to improve the instrument.


©1994 Reflective Thinking Associates

At this point, it does not appear that the RTA documents a systematically higher or lower pattern of response for men or women. In this respect, the RTA mirrors the general findings of the RJI in that gender differences, if they are present in the instrument, appear to reflect differences in educational attainment and not gender-related differences in preferred cognitive style. This interpretation must be cautioned by the small numbers of male graduate students assessed, and by the possibility that level of reflective judgment may affect a students choice to pursue or complete graduate study.

Although the RTA appears to document internally consistent differences in performance across educational levels, the relationship of the composite score of the RTA to the RJI is as yet unknown. Given the smaller effect size associated with educational level, it seems reasonable to assume that differences between individuals in their ability to produce solutions to ill-structured problems (as measured by the RJI) are more pronounced than individuals' abilities to recognize and choose between stated solutions (as measured by the RTA). Although the evidence gathered so far is promising, additional research which assesses individuals longitudinally (or at least assesses more carefully controlled homogeneous samples of students) is called for. Controlled studies which examine the roles of general verbal ability and gender are also necessary before performance on the RTA can be as confidently attributed to the effects of college education as the research to date using the RJI.


©1994 Reflective Thinking Associates

## References


Davison, M.L. (1979). Testing a metric unidimensional qualitative unfolding model for attitudinal or developmental data. Psychometrika, 44, 179-194.

Davison, M.L., King, P.M., Kitchener, K.S. & Parker, C.A. (1980). The stage sequence concept in cognitive social development. Developmental Psychology, 16, 121-131.

Rest, J. (1979). Development in judging moral issues. University of Minnesota Press: Minneapolis, MN.

Rest, J. (1979). Revised manual for the Defining Issues Test. Minnesota Moral Research Projects: Minneapolis, MN.

Wood, P.K. (1994) A secondary analysis of claims regarding the Reflective Judgment interview: Internal consistency, sequentiality and intra-individual differences in ill-structured problem solving. unpublished manuscript available from Phillip Wood, 210 McAlester Hall, University of Missouri-Columbia, Columbia, MO 65211

Wood, P.K. (1994b). The effects of unmeasured variables and their interactions on structural models. in A. von Eye and C. Clogg (Eds.) Latent Variable Modeling. Sage: Hillsdale NJ.




©1994 Reflective Thinking Associates

Figure 1: Overall RTA score by Educational Level (Based on ind.'s w/mean.<.05)

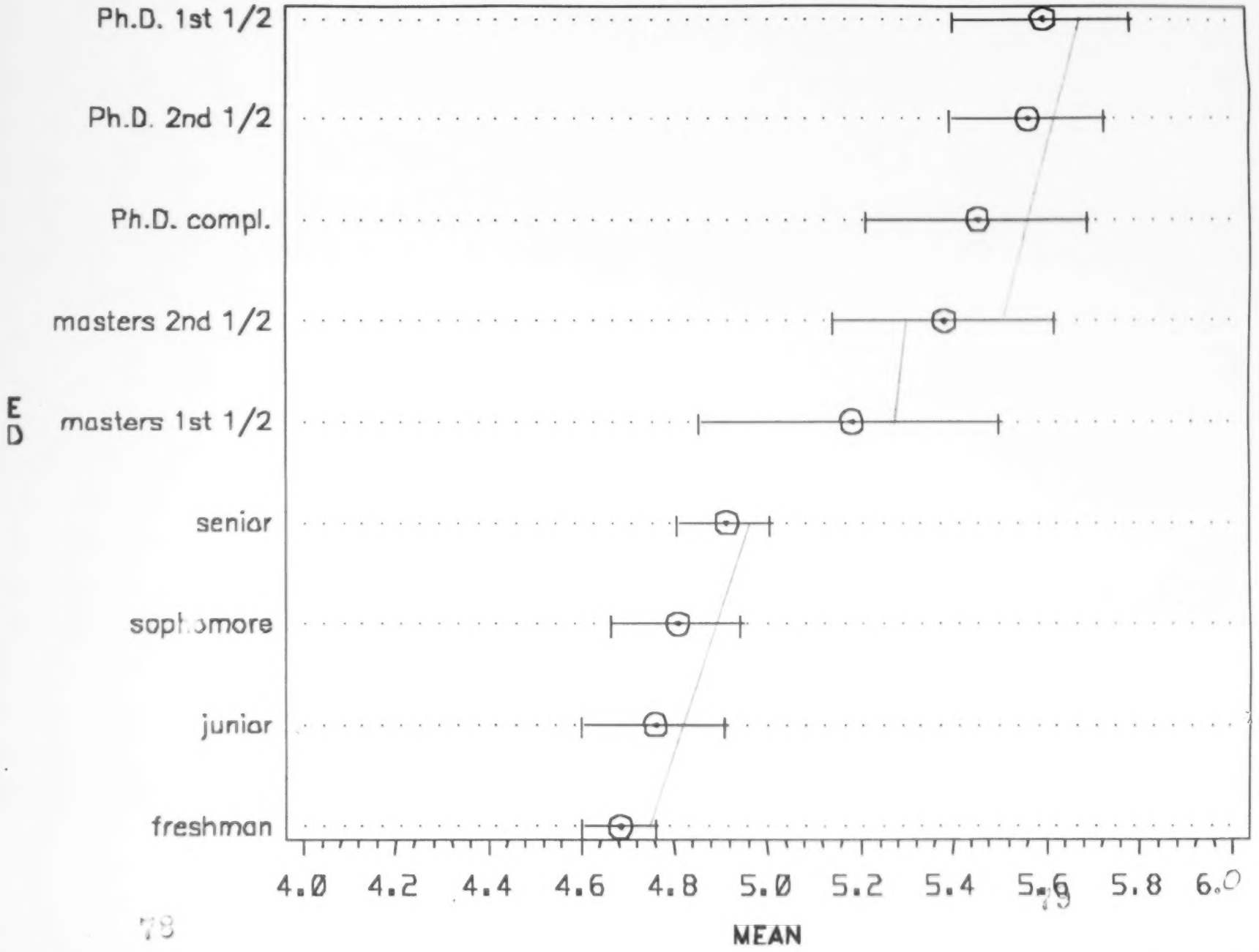
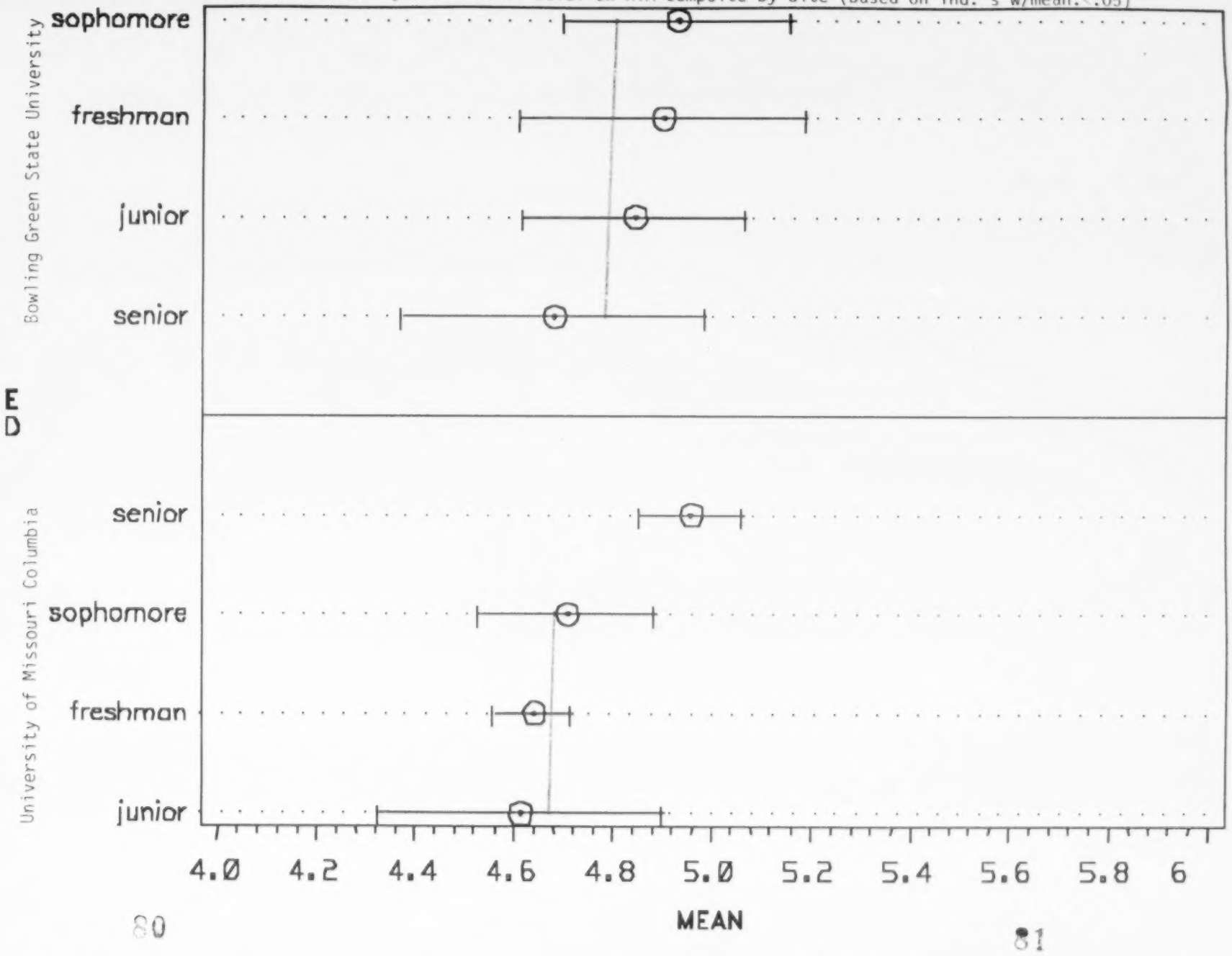


Figure 2: Means by Educational Level on RTA Compsite by Site (Based on ind.'s w/mean.<.05)



APPENDIX C

**EVALUATION**
**DEVELOPING REFLECTIVE THINKING IN THE COLLEGE CLASSROOM**
**JUNE 20, 1992**

(please feel free to continue your comments on the back of this page)

1. In general, the format of today's meeting was: (please circle)

   poor adequate good very good

2. The format of today's meeting could have been improved by:

3. The mini-lecture on the role of feedback and support in promoting Reflective Judgment was: (please circle a number)

   useless 1 2 3 4 5 helpful

   confusing 1 2 3 4 5 clear

   Comments:

4. The mini-lecture and demonstration on teaching techniques was:

   useless 1 2 3 4 5 helpful

   confusing 1 2 3 4 5 clear

   Comments:

5. Comments on group discussions:

6. Overall, the information I received in this workshop was:

   useless 1 2 3 4 5 helpful

   confusing 1 2 3 4 5 clear

   Comments:

7. The application portion of this workshop could be improved by: (If you say "more time," please indicate what you would omit if more time were not available.)

8. Other comments:



BEST COPY AVAILABLE